EXPERIMENTATION and MEASUREMENT in RELIGIOUS EDUCATION

GOODWIN B. WATSON, Ph.D.
Assistant Professor of Education, Teachers College, Columbia University
Director of Research, Home Division, National Council
of the Y M C A

ASSOCIATION PRESS
NEW YORK: 347 MADISON AVENUE
1927

COPYRIGHT, 1927, BY
THE GENERAL BOARD OF
THE YOUNG MEN'S CHRISTIAN ASSOCIATION

PRINTED IN UNITED STATES OF AMERICA

TABLE OF CONTENTS

List of Tables viii

Introduction xi

CHAPTER PAGE

I. The Scientific Approach to Religious Education 1
1. Magic *versus* Scientific Method 1
2. Magic and Method in Religious Work 2
3. Problems Not Answered by Experiment 3
4. Experiments Needed Today 4
5. Pure Research and Practical Research 5
6. Sources of Problems for Practical Research 6

II. The Choice of Experimental Method 10
1. Principles Governing Choice of Problem 10
2. Schemes for Organizing Experiments 13
a. Single-Group Methods
b. Equated-Group Methods
c. Rotation Methods
d. Survey Methods
3. General Advices 30
a. Miniature Experiments Advisable
b. The Diary
c. Experiments with Individuals

III. Methods of Measurement 34
1. Observation as Measurement 34
a. Whatever Exists Can Be Measured
b. Methods of Refining Crude Observation
c. Value of Ratings
d. Reliability of Measures
e. Validity of Measures
2. Objective Records as Measures 53
a. Methods of Collecting Objective Data
3. Paper and Pencil Tests 57
a. Tests *versus* Questionnaires
b. Ethical Problems in Testing
4. Conduct Tests 59
5. Interview Methods 61
6. Case Studies 62

IV. Tests Now Available 67
1. The Growing Field 67
2. Intelligence and Vocational Tests 68

268
W333e

80720

CHAPTER PAGE

3. Description of Tests Now Available 70
   - Biblical Knowledge Test (Whitley)
   - Bogardus Social Distance Test
   - Brief Test in Religious Education
   - Brotmarkle Comparison Test
   - Church School Examination Alpha
   - Colgate Emotional Hygiene Test
   - Downey Will-temperament Test
   - Emotional History Record
   - Fernald Achievement Capacity Test
   - Fernald Ethical Discrimination Test 36,035
   - Fernald Ethical Perception Test 27,105
   - Freyd's Occupational Interest Blank
   - Giles, Sunday School Examination A
   - Hart Personnel Assayer
   - Hart Test of Social Attitudes and Interests
   - Interest Analysis
   - Kent Rosanoff Free Association Test
   - Koh's Ethical Discrimination Test
   - Laycock's Test of Biblical Information
   - Lundholm, Emotional Cross-out Test
   - Miner's Analysis of Work Interest Blank
   - Multiple-choice Test of Religious Ideas
   - Pressey X–O Tests for Investigating the Emotions
   - Racial Attitudes Test
   - Social Relations Test
   - Survey of Public Opinion on Some Religious and Economic Issues (Watson Prejudice Test)
   - Union Test of Ethical Judgment
   - Union Test of Religious Ideas
   - Upton Chassell Citizenship Scale
   - Woodworth Matthews Questionnaire

4. Valuable Tests Not Now Available 97
   - *a.* Case's A True-false Test in Religious Education
   - *b.* Chapman's A Test of Motives
   - *c.* Character Education Inquiry Tests
   - *d.* Chassell Parable Interpretation Test
   - *e.* Fahs Biblical Test
   - *f.* Fernald, Meritorious Acts
   - *g.* Hartshorne, Tracy, Unfinished Stories
   - *h.* Self-ordinary Ideal Rating Scale
   - *i.* Laslett, Controlled Association Test
   - *j.* Orr, Good Manners Test
   - *k.* Porter Advanced Bible Test
   - *l.* Porter, Student Opinion on War
   - *m.* Raubenheimer Character Preference Test
   - *n.* Schwesinger Socio-ethical Vocabulary Test

CHAPTER PAGE

o. Travis, Test of Personality Traits
p. Van Wagenen Character Judgment Scale in American History
q. Y M C A Religious Education and Character Growth Tests

5. Life Situation, Conduct, and Behavior Tests 106
a. Aggressiveness
b. Caution
c. Cheerfulness
d. Civic Duty
e. Concentration
f. Confidence
g. Conformity
h. Courtesy
i. Decision
j. Delinquency
k. Emotionality
l. Group Loyalty
m. Helpful Behavior
n. Home Environment
o. Honesty
p. Honest Confession
q. Humor
r. Imagination
s. Interest
t. Negativism
u. Persistence
v. Recklessness
w. Sociability
x. Social Perception
y. Studiousness
z. Suggestibility

6. Use of Standardized Tests 122
a. Discussion of Test Results
b. Cooperation in Standardization

V. The Construction of Tests 126
1. Definition of What Is to Be Measured 127
2. Selection of Content 127
3. Form for Test Elements 128
a. New Type and Essay Type
b. True-false
c. Two-answer
d. Completion
e. Word-phrase Answer
f. Multiple Choice
g. Degrees
h. Ranking

CHAPTER PAGE

i. Matching and Pairing
4. Other Significant Classifications 145
5. Development of Directions 148
6. Methods of Determining Scoring Scheme 150
a. Scoring When Best Answers are Known
b. Scoring on Basis of General Criterion
c. Technique of Criterion Score Process
d. Weighting Test Elements
7. Standardization of Tests 159
a. Objectivity
b. Reliability
c. Validity
d. Comparability
8. Criteria for a Good Test 166

VI. Statistical Methods 171
1. The Service of Statistics 171
2. Distribution of Scores 176
a. Order
b. Rank
c. Frequency
d. Graph
3. Measures of Central Tendency 183
a. Mean
b. Median
c. Mode
4. Other Significant Points 187
a. High Quartile
b. Low Quartile
c. Deciles
d. Percentiles
e. Marks
5. Measures of Variability 191
a. Range
b. Average Deviation
c. Standard Deviation
d. Q
6. Measures of Relationship 197
a. Graphic Measures
b. Correlation
c. Regression
d. Critical Score
e. Partial Correlation
f. Partial Regression
g. Multiple Correlation
7. Practical Problem in Five Variables 224
8. Measures of Reliability 237
a. S. D. and P. E.

CHAPTER PAGE

b. Reliability of Prediction from a Regression Equation
c. Reliability of a Difference
d. Limitations of Reliability

VII. Presentation of the Results of Experimentation 245
1. Choice of Audience 245
2. Written Reports 247
3. Use for Publicity 248
4. Education of Those who Participate 251
5. Twenty-five Pitfalls 252
6. Religion and Scientific Research 254
Appendix 256
Suggested Problems for Experimental Investigation 256
The Construction of a Sample Scale 270
Statistical Tables 273
Glossary 279
Index 292

LIST OF TABLES

TABLE NO.		PAGE
I.	How to Pair Subjects for Equated Groups	21
II.	Suggested Form for Graphic Rating Scale for Initiative	41
III.	Form for the Weighting of Test Elements in Accord with Criterion Scores	157
IV.	Order Distributions: Showing Also Mean, Median, Mode Q_1, Q_3, Q, and Range	177
V.	Rank Distributions	179
VI.	Frequency Distributions: Showing Also Median, Q_1, Q_3, and Q	180
VII.	Graph Distributions	182
VIII.	A Curve of Normal Distribution	183
IX.	Average Deviation and Standard Deviation	192
X.	Use of Standard Deviation in Comparing Scores	195
XI.	Graph of Relation between Ethical and Religious Scores	198
XII.	Rank Method Correlation between Intelligence and Ethical Score	200
XIII.	Rank Method Correlation between Intelligence and Religious Score	201
XIV.	Rank Method Correlation between Ethical Score and Religious Score	202
XV.	Correlation between Intelligence and Ethical Score	204
XVI.	Correlation between Intelligence and Religious Score	205
XVII.	Correlation between Ethical Score and Religous Score	206
XVIII.	Example of Correction for Attenuation	210
XIX.	Prediction of the Most Probable Ethical Score from a Given Intelligence or Religious Score	211
XX.	Partial Correlation between Ethical Score and Religious Score When Intelligence is Held Constant	214
XXI.	Summary of Data for Hypothetical Problem in Administration of a Board of Religious Education	226
XXII.	Intercorrelations of Measures Obtained by Board	226
XXIII.	The Reliability of Certain Measures Previously Computed	239
XXIV.	The Reliability of the Predictions in Table XIX	241
XXV.	How the Board of Religious Education Should Spend Each Dollar	250
XXVI.	Form of Recording the Number of Judges Preferring One Specimen to Another, for Ten Samples	271

LIST OF TABLES

TABLE NO.		PAGE
XXVII.	Conversion of Per cent of Judges Believing a Given Sample Superior to a Second Sample, into the Distance in S. D. Units between First and Second Samples	272
XXVIII.	Squares and Square Roots of Numbers to 100	273
XXIX.	Value of r Corresponding to Each Value of rho	274
XXX.	Predictive Indices for Certain Correlation Coefficients	275
XXXI.	Interpretation of Correlation in Terms of Displacement	276
XXXII.	Per cent of Independent Causal Factors Common to Two Measures Yielding Certain Correlations	277
XXXIII.	Per cent of Total Number of Cases (Area of Normal Curve) Falling above Given Scores in S. D. Units	278

INTRODUCTION

This presentation is intended for students in undergraduate and graduate courses in religious education who are anxious to secure a more adequate understanding of the experimental viewpoint and of the necessary techniques. It is hoped that it may prove useful, as well, to enterprising pastors, directors of religious education, secretaries of the Y M C A or Y W C A, and similar professional workers in the field of religious education. Officials charged with the supervision of large areas, writers of curricula, and persons responsible for administrative policy may find suggestions which will help them to establish a new method of work.

The needs of so varied a group demand a text which shall contain some explanations unnecessary for advanced students, and some techniques too complicated for the service of the average experimenter. The author hopes that readers will excuse his errors of judgment in selection and organization with the thought that somewhere in so diverse a population there may be those whose needs will be served.

Little is offered which is original, save the applications to the complex problems of character development and religious education. Experimental forms, methods of scale construction and test making, and statistical methods may all be found discussed more fully in other volumes. It has been the attempt of the author to bring together in one volume for convenient study and reference by the experimenter in religious education the most essential materials, previously scattered through periodicals and dozens of books.

The reader inexperienced in the forms and formulae here presented may find it advantageous to read through the entire book first, passing over the points which are not clear. Later he may wish to read it again more carefully, working forward

and backward, using index and glossary, and working out the formulae and statistical processes. The first survey will perhaps open up the field of possibilities in experimentation and measurement, and lead to awareness of techniques which need to be acquired. The book is so written, however, as not only to serve this need but to form a handbook of reference for the more experienced student.

Fortunately, this book will often be used by those who are actually carrying forward experiments. It can do them little harm. Having once tasted the thrill of the experimental establishment of previously undiscovered truth, no presentation of schemes and formulae, however inept, will long deter them from further indulgence. The sharp barb of curiosity, the patient eagerness of search, the culminating mastery of accomplishment speak with a potency which words will not convey.

It is the author's privilege to acknowledge indebtedness to Professor George Albert Coe for stimulating viewpoints which are implicit in the philosophy of religion here assumed. To Thorndike, Garrett, Otis, and particularly to William A. McCall the author is anxious to give credit for all the best statistical formulae, tables, and explanations. To his associate, Ralph B. Spence, and to his wife, Gladys H. Watson, are due thanks for reading the manuscript and proof, and for numerous useful suggestions.

GOODWIN B. WATSON

New York City,
July 1, 1926

CHAPTER I

THE SCIENTIFIC APPROACH TO RELIGIOUS EDUCATION

1. Magic versus Scientific Method

Human progress has involved as one of its essential features, the substitution of scientist for soothsayer. In the primitive village sickness was treated with magic ointment and supposedly potent incantations. Messages were sent to the absent by means of favorable spirits, invoked by suitable shibboleths. Ceremonies and spells were accepted without blasphemous questions. An impertinent inquirer possessing the temerity to ask, "Why?" "How do you know?" and, "Will it always work that way?" might expect uncomfortable consequences after the devout had recovered from their astonishment.

In modern civilization sickness is carefully studied, the symptoms diagnosed, and remedies prescribed. Messages are sent over wires, or indeed without wires at all. The scientific student of medicine or radio transmission is not dismayed by those who would ask "Why?" or "How do you know?" or "Will it always work that way?" Underneath all his activity is an assumption that the most carefully controlled experiments with most accurate measurements will yield truth which is wholly reliable. In the control of the physical universe magic has been widely replaced by method. Rivers are spanned, skyscrapers arise, light and power and thought are transmitted with incredible swiftness. Somewhat more slowly the processes of education and of economic life have emerged from the domain of blind, self-satisfied prejudice into the keen cool atmosphere of the scientific search for laws which will always hold true. Equations are being written which will express the future increase in a city's population, the probable consequences of a bankruptcy, the total energy available in a

human being with a certain diet, the rate at which he can learn to read and write, or the rate at which his mind will change under the application of given stimuli. Achievement has been the reward of open-mindedness, humility, and patience. The will of God in the life of the universe has been found written clearly for those willing to strip aside the placards of tradition and to search reverently for the relationships which constitute reality.

2. Magic and Method in Religious Work

Religion is still apt to be frightened by the imperious scientific search for truth. Suppose some impertinent inquirer to have escaped after his experience in questioning the primitive practices of the medicine men. Let him approach a certain Sunday school and ask, "Why do you sing hymns? What evidence have you that those prayers and sermons will make these varied children into what you would like to have them become? Why do you study such complicated material as the Scriptures? What are the actual consequences of your methods? Do they always work? Why not give up the whole business? Would not society be better off?"

Sometimes the answers would be emphatic opinions of workers who have grown up under such methods and believe in them; frequently the answers would point to the fact that the practice is very old or very widespread. Such answers would differ little from those made by the tribesmen. Indeed, it is not at all impossible that the inquirer would find, after the officials recovered from their astonishment, that he was no more welcome in their ceremonies than he had been among the savages.

Today this stronghold of the soothsayer is crumbling before the attack of a scientific spirit in a new generation. Pointed queries as to *whys* and *wherefores* are met with more investigation and experiment, and with less of that emotional reaction which indicates a lack of intellectual defenses. An ever-increasing number of workers in religious education are more anxious to find truth than to confirm their previous opinions or blindly to preserve the traditions of the past. Frequently, of course, they find that race experience has reached the goal before them, that the methods, institutions, and viewpoints which have been un-

critically accepted have had more than a large nucleus of truth. It is only reasonable to expect that any system of thought or action which has commended itself to any large group of people for a considerable time, would have in it elements not lightly to be tossed aside. This expectation is as applicable to Fundamentalists as to Modernists, to Buddhists, Mohammedans, Bolsheviks, Rotarians, Pacifists, Militarists, the Ku Klux Klan, and the Knights of Columbus, as to the particular viewpoints any one group may find most congenial. Therefore, the scientific approach cannot rest satisfied with the assumption that there is something worth while in a practice because some people think so or thought so. The student of the ways of God may not assume in advance of his investigation which of the elements of tradition are true and valuable. Such findings may be his conclusion but a broad skepticism must be his starting point. True, the scientist in the field of religious education is not without assumptions. He assumes that God will reveal himself to those who seek in the manner in which God has chosen to be sought; that God is reliable rather than whimsical; that the most uninviting truth is better than the most heart-warming error; and that only by painstaking exactness can he hope to approximate to the real nature of the complicated relationships with which he deals.

3. Problems which Cannot be Answered by Experiment

It is ever the fault of those who feel strongly the need for some point of view which has not been stressed by their fellowmen previously, that these pioneers become overzealous in the cause of their new ideas. Scientists have not always spoken to laymen with the same humility which would characterize these scientists in relation to the physical universe. Striplings in the world of ideas are apt to show a jaunty self-confidence which older and better established viewpoints do not need.

It would be unfortunate, but not surprising, if the scientific approach to problems of religious development were to claim more than it can fulfill. Because so much will be said about problems which can be best answered by experimentation it may be well to emphasize the fact that some of the most important

questions of life cannot be settled in that way. Consider the following problems.

1. Do the people of Tennessee have a moral right to prohibit the teaching, to coming generations, of doctrines which are repugnant to the present generation?
2. What is truth? Reality? Objectivity?
3. What is beautiful?
4. What sort of life did Jesus live? What would he do in such a problem as mine?
5. Is militarism preferable to pacifism?
6. Do all men have an equal right to happiness?
7. Should a few suffer for the good of all?
8. Is western civilization with its rapid progress more satisfactory than the quieter meditations of oriental countries?
9. What is the aim of religious education?

Upon all such questions scientific experimentation may contribute data. None of them is completely answered when the report of the scientist has been made. Indeed in any matter involving human welfare the scientist can state only, "If this is done — these will be the consequences." Whether people choose to take one set of consequences or another rests eventually upon a value judgment or preference which goes beyond science, however much use it may make of scientific processes and findings.

4. Experiments Needed Today

With due recognition of limitations, the impression still remains of the overwhelming need for accurate appraisal of the teachings and teaching of religion. In the appendix (pp. 256) are suggested a hundred problems many of which are crying for experimental investigation today. Upon most of them, practice is proceeding blindly. To some of them, such, for example, as the consequences of prayer for the sick, many people would feel the answer established. Yet none of them have had adequate scientific investigation. All of them are questions of fact which depend for answer upon the results of actual experiment, or the gathering up of evidence from the experiments which are constantly being performed in the natural course of living. None of them are capable of reliable solution by armchair or dialectic

methods. No decrees upon any of these questions, whether set forth by curriculum makers, church leaders, professors of religious education, or experienced church school workers are worthy of acceptance, except insofar as they rest upon careful descriptions of what has actually happened. Any statement upon any such problems at the present time should be accompanied by due consideration of the limits within which actual observations have been made and of the crudity of the measures of appraisal.

5. Pure Research and Practical Research

Faced by such a stupendous task, by questions which may not be answered for generations, the practical worker in religious education may well pause. What is to be done? Are children and young people to be left to their own resources while the army of religious workers turns with one accord to test tube and stop watch and statistical graph? Surely in some way business must be carried on at the old stand, while the more glorious building is being designed.

There are some workers who should be set free, and adequately equipped to carry forward tasks of pure research. They should be encouraged to give their lives to the exact and laborious search for the true answer to a detailed problem. By this method every firm coral island of science has been built. Out of such investigation emerges eventually an unambiguous answer to the chosen question, based upon evidence which any competent mind finds adequate. No restrictions or shackles should be placed upon such workers, save that they shall be rigorously scientific, impartial, thorough, and accurate. They should be protected from influence by men of affairs who are "practical" in a hard and limited fashion. Thorndike has well called the attention of educators to two great men who lived in England during the same generation. One was Thomas Arnold, who gave his life to the significant task of improving actual schoolroom practice. He stands out today as a great example in the quickly neglected past. But Francis Galton, who may never have taught in a single schoolroom, working in the realm of pure research, achieved results which permeate the theory and practice of almost every modern educational system.

An equally great contribution to the progress of scientific religious education can, and should, be made by the teachers and administrators who are linked up with practical field situations. The greatest present hope for experimental results rests upon them. Their own advance in theory and improvement in technique depends in large degree upon their success as experimenters. It is to be expected that such workers will not launch out on any problem merely because it strikes their fancy. They will carefully analyze the work they are carrying on. They will find the points at which they are failing to achieve maximum results. They will select among possible devices the few which seem most likely to be successful. These will be put into operation. The success of the experiment will necessarily depend not so much upon the securing of a final theoretical answer as upon the practical progress of the program. A solution which is all right in theory but which does not work in practice, is no solution at all from this point of view.

6. Sources of Problems for Practical Research

Experiments of this second sort are being performed constantly by every alert teacher and administrator. The worship service doesn't seem to go well. A song leader is introduced. Improvement follows. James insists upon disturbing the group. The teacher calls at his home. He appears worse the next week. He is sent to the superintendent. He comes back cowed, but takes no further constructive part in the class work. So it goes. Sometimes with groups and sometimes with individuals experiments are constantly under way. For the most part, however, they are measured only by the uncritical lump judgment of one teacher or administrator. Frequently the factors producing the change are so muddled together in a confused situation, that what seems to be effective in one case appears not to work in another. By careful control of the factors which might influence change, the investigator may discover the "key" to the situation. If he succeeds, he has found an operation which, under given conditions, will always and invariably produce given results. That this is not apt to be the result of the ordinary trial and error progress of the field worker is due in large degree to careless-

ness in definition of problem, failure to isolate operations employed, and crudity of the measures applied to the subjects of the experiment.

Not only in these day-by-day experiences with new devices, new techniques, and new subject matter in his own field situation may the practical research worker find his problems. A most fruitful field of inquiry for experimenters is indicated by the heated controversies among other workers. Dr. A. advises hand work for primary children, Dr. B. condemns it as a waste of time so far as attitudes are concerned. The Y M C A in one section of the country may desire a program of medals, awards, and standards, while in another section, equally able leaders insist that each club should work out its own program, that initiative and local adaptability require more flexibility than such a standardized program can give. Some teachers believe that lectures are efficient, others insist upon discussions only, while still others believe a combination secures the most desirable learning. So it goes with one issue after another. Each party to the controversy produces illustrations of the favored point of view. Occasional experiments are made, tending usually to confirm the faith of the experimenter. The opponents point out the inadequacy of these experiments and produce others. Why not substitute a more sensible procedure? Why not get together, in advance, the parties to a controversy? Why not get them to plan cooperatively a research enterprise which they would all expect to yield significant results? It would seem that real progress at many points might be achieved by taking into account, before the experiment is made, all the people who are likely to be called upon to make use of its results. Usually this will lead to modifications which make the experiment of much more worth. At any rate, it will lead to an attitude on the part of the teachers and leaders likely to make for a surer integration of the experimental results in their practical work.

A leader in organized religious movements in the United States questioned the value of such cooperative setting-up of practical research, because of the prejudices which are truth proof. "It might be possible," said he, "to get scientists who thoroughly

respect one another to agree upon such a procedure; it might be possible to get religious workers to submit minor points of difference to such tests; but the heated controversies involve so much of emotional habit, that no evidence would suffice to bring about a change in viewpoint on the part of the religious leaders." There is enough probability that he is correct, to cause grave concern to any interested in progress toward truth about religious education. It seems more encouraging and quite as reasonable to believe that there is a great body of "in-betweens," many of them eager for real light on the difficulties involved in character formation, who would welcome the opportunity to participate in planning an investigation and in utilizing its findings. The advance of the science of religious education rests in large degree upon the development, in the rank and file of workers as well as in the leaders, of an expectation that controversies are not to be settled by majority vote, partisan loyalty, or impassioned oratory, but by careful, painstaking endeavors to get the facts

EXERCISES

1. Consider the following statements. If one seems to you based primarily upon "magic" place an *M* in the parentheses following it. If it seems to be a result of careful observation of what happens under given circumstances, and to be more scientific than magical, place an *S* in the parentheses.

a. The night air makes people ill..... ( ).
b. Reading Psalms brings comfort..... ( ).
c. Children hearing about Daniel become braver..... ( ).
d. Thirteen is unlucky..... ( ).
e. Prenatal shocks mark a child..... ( ).
f. Communication across the ocean without wires is possible ( ).
g. The scientist has authority which is to be trusted..... ( ).
h. Christians are happier than Mohammedans..... ( ).

2. Which of the above statements are capable of scientific investigation?

3. Consider a particular religious activity, such as a prayer meeting, a conference, or a week-day class. Follow through the steps outlined for practical research in such a field, indicating what sort of problem might arise, and when it could be considered solved.

4. Rank the following investigations in the order of their probable contribution to the progress of the world as viewed one hundred years from now. Put 1 in front of the one you would expect to be most influen-

tial, 2 next, then 3, 4, and finally 5 in front of the one which you believe will seem then most trivial. Justify your answers.

____ *a.* What different ways of handling what situations in childhood would lead to the employment of cooperation rather than attack and defense attitudes?

____ *b.* What are the most useful forms of report cards for church schools?

____ *c.* Under what conditions may Modernists and Fundamentalists be led into mutual accord and progress?

____ *d.* Does an amoeba think?

____ *e.* Can acquired characteristics be inherited?

5. List five important problems of life upon which scientific study can throw no light. Do not include those mentioned in the chapter.

6. Mark each of the following statements by placing after it a figure from this scale: Use 5 to mean surely completely true; 4 to mean probably true, or true in large degree; 3 to mean doubtful, unknown, about even; 2 to mean probably false, or false in large degree; 1 to mean surely completely false.

a. General truth is more needed than practical solutions... ( ).
b. Whatever is true is important..... ( ).
c. Spiritual growth follows predictable paths..... ( ).
d. The finding of natural causes means the depreciation of belief and trust in God..... ( ).
e. Scientific research is prayer..... ( ).
f. Emotional defenses usually arise in the absence of adequate intellectual defense..... ( ).
g. Common sense has preceded science in finding the truth about all important human relationships..... ( ).
h. Most of the wasted effort of men in religious work is due to the uncritical acceptance of errors about the way in which things happen..... ( ).
i. A preexistent bias is likely to determine the kind of results obtained in an experiment..... ( ).
j. Some things are too sacred to investigate..... ( ).

CHAPTER II

THE CHOICE OF EXPERIMENTAL METHOD

Students of the processes of religious education who find themselves anxious to participate in a scientific search for the laws which describe the true relationships of life, often find their work ineffective because of unwise demarcation of their specific problem, and of ill-adapted methods of investigation. Let us suppose that a pastor in a progressive church, or a student entering upon his Ph.D. work, is anxious to make an experimental contribution. Suppose he has considered a wide range of suggested research problems. Suppose he has weighed the merits and disadvantages of a study in pure research as opposed to one of practical research. If interested in the latter, he has studied particularly his own field of work, or the points of controversy in which practical workers are involved. There are several principles which he should bear in mind as he chooses from this array the problem to which he is to give his effort.

1. Principles Governing the Choice of a Problem

a. Other things being equal, the problem should be one of vital concern to many people. There are too many problems, and too few investigators, to make it much less than deplorable waste for any man's trained skill to be expended upon problems which can only by utmost benevolence be regarded as contributions to humanity. Of course this principle must not violate the right of the worker in pure research to chart the field as he sees it, and to select as most vital, problems which contribute to remote rather than immediate ends.

b. Other things being equal, the problem should be one in which the investigator is personally interested. There are hours of painful detail in almost every research worth making. A question which arouses only half-hearted concern at the beginning of a study may sometimes become more and more intriguing as

the work progresses, but too often will the task seem more and more tedious, and efficiency will wane with interest.

c. Other things being equal, the problem should be such that the investigator is equally happy, whatever the true result. Few experimental procedures are completely free from opportunity for the unconscious bias of the investigator to exert its influence. The attacks and counter-attacks launched by people representing varying aspects of a heated controversy, against the experimental work of their opponents, are not due entirely to the refusal of the speaker to accept scientific evidence. They are sometimes due to the too well grounded suspicion that somewhere in the process the concern of the investigator has found unconscious expression. Intellectual, scientific detachment from emotional responses may be possible for a few people. The safer rule is to choose a question in which the investigator's interest is in the question, not in one or the other of the proposed solutions.

d. Other things being equal, the problem should be one which has not previously received adequate investigation. It is a specious and self-defeating originality which would attack problems in ignorance of previous studies in the same field. "Not necessity, but knowledge of other men's inventions," says Thorndike, "is the mother of invention." Of course this principle should not prevent the repetition of experiments which are regarded as of crucial importance and which have not been adequately checked. In the entire field of psychology and education, experimental progress has come so fast that there has been inadequate testing of results by repeated experiments. Each man has felt the urge of problems not yet investigated at all. Duplication has seemed unduly costly. Yet it is probably safe to predict that some of the most significant changes in practice among educators 50 years from now will be due to the overthrow of conclusions now widely accepted on the basis of meager and unrepeated experiments. In the field of religious education it is probable that the greater need at present is for the study, based on scientifically established data, of problems which have as yet no valid solutions. There is little as yet to be repeated and checked.

e. Other things being equal, the problem should be one adapted

to the materials and persons available. In some situations, intensive study of a few children is indicated, in others, a similar problem should be investigated by broad survey methods because of the contacts which the investigator has at his disposal. Not only persons but finances are often limited. Research is costly. Religious institutions are probably due for severe shocks when they find the expense involved in answering some of the simplest of their questions. A committee recently estimated that to get an approximate answer to the question of the comparative value of systematic Bible study and modern life-problem discussions in the Hi-Y clubs of New York state would cost $5,000, exclusive of the services of the secretaries already employed and able to devote a portion of their time to the work. Of course such expenditures may prove to be exceedingly economical in the long run. There is nothing more costly than ignorance.

f. Other things being equal, the problem should be so sharply defined that it becomes relatively easy to separate relevant from irrelevant inquiries. It is probably impossible to foresee all of the false leads in the first investigation of any problem. Yet to start out, as do many students interested in dissertations, planning to work a little at this and a little at that, intending that somehow out of the maze a problem shall emerge, is almost surely to waste time, money, and effort. Every experiment has its fascinating fringe. It would be interesting to find out this and that, or perhaps the other might be added as a side line. Experience has usually indicated that the side lines are not worth the effort. The experiment should be planned, as a rule, to give help with one definite problem. Other things should be ruthlessly eliminated. The omnibus experiment is a sure symptom of the amateur.

g. Other things being equal, the problem should be so limited that it can be finished within the available time. A student recently proposed for his thesis the problem, "A Study of the Effects of Four Years in College on the Character of Students." He proposed to investigate young people in and out of college with reference to all the factors involved in character. While the problem is undoubtedly important, it is probable that the

development of measures, and their application to pupils before college, then to students in all sorts of colleges, with adequate chance for follow-up in life should not be attempted in less than 20 years. It is a sophisticated worker who expects that an investigation will take him about one-third again as long as he expected it to take.

2. Schemes for Organizing Experiments

Given the specific problem, well adapted to the person making the investigation, the next task is the choice of suitable experimental methods. Many problems might be investigated by several different methods, but not often can a question be as well investigated by one method as by another. Experimental methods may well be studied with reference to four types: single-group methods, equated-group methods, rotation methods, and survey methods.[1] In each of these types there may be one, two, three, or any number of operating factors which are employed.

a. Single-group Methods

The simplest method is that which employs a single group. The everyday experiments of the religious worker belong to this type. Suppose there is a budget to be raised. The pastor may try the effect of a sermon on finances, followed by a collection. In such case the initial test is the number of dollars on hand — none, perhaps. The operating factor is the sermon. The results are recorded by the final test, *i.e.*, the number of dollars collected. It may be diagrammed as follows:

$$T_1\text{—Sermon}\longrightarrow T_2$$

This becomes more complicated when the experimenter employs several operations instead of only one. Suppose the first attempt to be relatively unsuccessful. The pastor may then try the effect of a letter to each member, then of personal calls, and finally of a big banquet. In that case, although the group is still just his congregation, there will be four operating factors. Diagrammed, such an experiment would be:

$$T_1\text{—Sermon}\longrightarrow T_2\text{—Letters}\longrightarrow T_3\text{—Calls}\longrightarrow T_4\text{—Banquet}\longrightarrow T_5$$

[1] An excellent presentation of schemes for organizing experiments may be found in McCall, "How to Experiment in Education," The Macmillan Co., New York, 1923, and in *Bulletin*, 1926, No. 24, U. S. Bureau of Education.

The tests represent, of course, the state of the budget at the end of each operation.

In another field, the single-group method is employed by the teacher who is anxious to select the best curriculum for his class in Sunday school. The initial test is his judgment that the sessions are rather dull, supplemented perhaps by comments from a supervisor or one or two pupils. He then introduces a new text. His final test may be his judgment that the class now shows more interest, but are still not inclined to study outside of class period.

Thus: T_1—New Text—T_2

This also becomes often an experiment with multiple operations. The first text succeeding moderately, another is tried but found to be worse, so a third is introduced, but finally the group returns to the first. The measures are, of course, very crude and unreliable, but upon them the progress of most groups now depends.

The single-group type may be illustrated as adapted not only to these crude enterprises, but also to refined techniques. Thus a student who wished to study the relative value of stories dealing with adults and stories dealing with children, very carefully equated six "adult" stories with six "children" stories. Tests were devised to measure the extent to which children remembered the main theme of the story. Only one group was available for experimentation. The stories were told to this group in alternation, so that the effect of time on retention might be practically equated for the two types. (Note that one type could not be used in position 1, 3, 5, etc., while the other was used in position 2, 4, 6, etc., because that would place the second group one period later, on the average.) The stories were thus told, one week apart, over a period of 12 weeks, with a single test at the end. The assumption, perhaps not wholly justified, was that the score for the group was zero on all this material at the beginning. Let us indicate Operation 1 (stories about adults) by O_1 and Operation 2 (stories about children) by O_2. Then it could be diagrammed:

T_0—O_1—O_2—O_2—O_1—O_1—O_2—O_2—O_1—O_1—O_2—O_2—O_1— Final Test.

The unpublished studies of Sturges in the field of attitudes are of similar type, except that the various operations consisted in repetitions of the same material. A group was tested with reference to their attitude toward pacifism and militarism, the test being so devised that 100 represented a complete, extreme militarist position, the scores decreasing until 0 represented a complete extreme pacifist position. Then Sturges presented Operation 1, which was 7 minutes spent in the silent reading of some printed material in favor of one end of the scale. The group were then retested on the same scale. This sometimes completed the experiment. Sometimes other operations were added. The group would be asked to read another 7 minutes, take the test again, read 7 minutes more and take it again, and finally to read a fourth period of 7 minutes, taking a final test. Diagrammed the experiment could be shown as follows:

$$T_1 \rightarrow O_1 \longrightarrow T_2 \rightarrow O_2 \longrightarrow T_3 \rightarrow O_3 \longrightarrow T_4 \rightarrow O_4 \longrightarrow T_5$$

In studying his results he wisely recognized that these operations were not entirely independent, and regarded them thus:

O_1 = Taking first test, plus reading 7 minutes.
O_2 = Taking first test, plus reading 7 minutes, plus taking second test, plus reading a second 7 minutes.
O_3 = Taking three tests, plus reading 21 minutes.
O_4 = Taking four tests, plus reading 28 minutes.

He then drew a curve to show the amount of change taking place from reading for 7 minutes, 14 minutes, etc. He found that people were more easily influenced at first. After the first period or two of reading, they had moved about as far as they were likely to move. It took much reading to bring about further change. He found that a mathematical equation could be written which would describe the way in which attitudes are moved when the average person reads the kind of material used in the experiment.

Single-group methods are acceptable under the following circumstances:

1. The purpose may not be one of comparison of methods or materials, but simply the statement that under given circum-

stances, such a factor produces such a result. In that case the usefulness is limited by the degree of completeness with which the operation is described and is reproducible, and by the representativeness of the group. If only one factor is involved, and it is simple, definite, and capable of being tried again by numbers of people, and the group is a fair sample of all the groups on which it might be desirable to try this operation, then the contribution from such an experiment is valuable. It is useful, for example, to know that given any group of babies, the application of a loud terrifying noise, together with a piece of candy, can be repeated to bring about a fear response when the candy alone is presented.

2. The purpose may be the study of cumulative effect. This necessitates study in one group. The study of learning, drill, or practice, means necessarily the repetition of operations on a single group. To be most useful, the group should be representative and the operation specific and reproducible, as above.

3. The purpose may be the comparison of factors such that having tried one, the possible influence of the second is not affected. Theoretically such situations are few, practically they arise occasionally. Thus in the study of the stories, there was not much carry-over from hearing stories about adults, which would influence the memory of stories about children, and *vice versa*. In the case of the budget, on the other hand, the limit may have been approached by some of the earlier operations. In that case, the operation tried at the end of the experiment would have little chance. The last thousand is the hardest to raise. Perhaps devices that seemed to fail then, would have succeeded beautifully if tried first or second.

The usefulness of the single-group type of method is obviously limited. Such experiments, although more common than those of any other type, have usually failed to make any widely useful contribution. It is very difficult to make the single group representative enough, and to find the need for comparison of operations which do not tend to affect one another. There is a further objection in the practical impossibility of "controls." Suppose it is made evident that after memorizing the One Hundred Twenty-first Psalm, a given class of boys were less

afraid of the dark than they had been when they entered school in the fall. Perhaps they simply outgrew the fear. Perhaps someone laughed at them, at home. Possibly the Psalm had nothing whatever to do with the apparent consequence. To secure adequate control, the consequences would have to be found to appear when the operation was tried, and not to appear when it was omitted from an otherwise identical experience. Only very rarely is it possible to use an operation which can be intermittently applied with little or no carry-over into the "control" period.

b. Equated-group Methods

The second general type of method employs equated groups. Thus, Houck and the writer, wishing to find out whether facts or stories which played upon the emotions were more useful in changing attitudes on the prohibition question, divided the total number of subjects into three groups on the basis of preliminary tests. To one group the experimenter read facts for 10 minutes. To the second group, alike in intelligence and in score on the prohibition question, he read Billy Sunday stories, for the same length of time. The third group, the control group, received no attention at all. Then all three groups took the initial test again. It was found that the first group moved 6 points (about 1 S.D.)[1] toward a dryer attitude, the second group moved almost exactly the same amount, while the third group averaged a slightly wetter score the second time than it did the first. It was concluded that the propaganda had definite influence, but that in these groups there was no appreciable difference between one type of propaganda and the other. His experiment might be diagrammed as follows:

Group I. T_1>—O_1(Facts)—→T_2
Group II. T_1>—O_2(Stories)→T_2
Group III. T_1>—Control——→T_2

Collings,[2] in his "An Experiment with the Project Curriculum," reports his plan for equating an experimental school

[1] See p. 193.

[2] COLLINGS, "An Experiment with the Project Curriculum," The Macmillan Co., New York, 1924.

district with two control schools. He endeavored to select districts alike in size, wealth, type of work done previously, social status, etc. Using teachers much alike in ability he carried on an experiment for 3 years in which the control schools did a good sort of old-type work, while his experimental school worked with great freedom along the lines marked out by the children's wholehearted purposes. His tests were necessarily many. He measured the actual achievement of pupils in reading, writing, arithmetic, and other standard subjects. He measured the social status of homes by modern improvements, cultural comforts, magazines read, community meetings held, etc. He found an indubitable superiority for his experimental school, as a result of the 3-year operation.

Gates[1] tried a similar experiment for 1 year with first-grade children, improving upon Collings' technique at several points. The children were more carefully paired to make the groups equal. The help given through supervision was made more nearly the same for both groups. The time spent by children on their school work was about the same in both groups. Gates tried to compare both "modern systematic" and "opportunistic" methods when at their best, rather than having one at average while trying to make the other superior. His results for only one year were not conclusive.

1. Cancellation of Variables in Large Groups

Sometimes the operations to be compared are tried out in a very wide range of groups, in the expectation that uncontrolled variables which might influence one group unduly, will be neutralized in the large variety of groups. Thus if on a chance division, half of the pupils in Methodist Sunday schools were selected to use a systematic course of study, while the other chance half were given a course of study which employed a "problem" approach and was arranged on a basis for "incidental" learning, measurement of those two groups, each containing some two million children, would undoubtedly yield valuable data. Probably the groups would be very evenly

[1] GATES, "A Modern Systematic *versus* Opportunistic Method of Teaching," Teachers College Record, pp. 679–700, New York, April, 1926.

equated. Poor teachers in one would be balanced by poor teachers in the other. Bright children in one would be, on the whole, offset by equally bright children in the other. All of this holds true, of course, only if the chance selection has been rigorous. If the division were on the basis of city schools using one curriculum and rural schools using the other, then many factors would make it hard to know just what the results meant. If geographical sections of the country formed the basis for division it might prove true that certain sections had advantages which others did not, thus unduly favoring them in the experiment. Particularly if the matter were left optional with the school, it might be expected that the results would be worthless for drawing conclusions about the curriculum. Perhaps the more modern, up-to-date, thorough-going school, with live, hard-working teachers would choose one type rather consistently. If so, then any superiority shown in the end could not fairly be said to be due to the course of study. If, however, the selection were made by going through an alphabetical list, taking every other name for one group and the remaining names for the other, or by drawing names out of a hat, then it is probable that any difference apparent at the end of the experiment could fairly be attributed to a difference in the value of the two courses of study.

When the groups are small, much more careful measurement is required. Usually it is necessary to "pair" each individual in one group with an equivalent individual in the others. General averages may not be enough. Certainly two groups are not paired if alike only in their average score. One may have many high scores and many low scores, while in the other they all cluster about the middle. This indicates the need for equating groups for "spread"[1] just as truly as for central tendency.[2]

2. Technique for Pairing

McCall has suggested a technique for pairing groups when a number of factors need to be taken into account. This, in general, is the plan.

(*a*) Measure every factor which might be significant. Thus,

[1] p. 191.
[2] p. 183.

for example, to study the effectiveness of the different methods of teaching college Bible classes, the groups would have to be equated for intelligence, age, Sunday-school training, present courses of study, interest in the subject, etc., as well as simply for Bible knowledge.

(*b*) Combine these measures into a single score for each individual. If they are simply added, each factor will be weighted in proportion to its standard deviation.[1] If some factors are believed to be much more important than others, they may be multiplied by two, three, four, etc., to bring them into proper influence. Thus for example, in the above-mentioned experiment on prohibition attitudes, the two factors of mental age and prohibition score were measured. It was found that the standard deviation of the mental age score was 11.5 while that of the prohibition attitude test was 6.2. If those scores had been added together for each individual, mental age would have been almost twice as significant as prohibition attitude in determining the position of one individual with reference to others. It was desired, however, to weight the attitude test twice as heavily as the intelligence test. Therefore, all the prohibition scores were multiplied by four and added to the mentality scores to make a composite score.

(*c*) Arrange all the subjects in order of their scores, highest at the top. Then distribute them as indicated in Table I. The first column is a possible series of composite scores arranged in order. The second column shows how the persons should be distributed in order to get two equivalent groups. The third column suggests the order for three groups. The fourth column suggests the order for four groups. The scores have not been divided by the naïve method of putting the first in one group, the second in the next, etc., because any such method leaves the first group with an average, on the whole, one step higher than that possessed by the group which contains numbers 2, 4, 6, 8, etc., from the list. In the suggested distributions, the averages remain practically the same for each group. Each high score tends to be balanced by one correspondingly low.

[1] See p. 193.

TABLE I. — HOW TO PAIR SUBJECTS FOR EQUATED GROUPS

Composite score	To obtain two groups	To obtain three groups	To obtain four groups
234	I	I	I
230	II	II	II
228	II	III	III
226	I	III	IV
225	I	II	IV
225	II	I	III
221	II	I	II
220	I	II	I
220	I	III	I
220	II	III	II
218	II	II	III
217	I	I	IV
217	I	I	IV
216	II	II	III
215	II	III	II
214	I	III	I
212	I	II	I
210	II	I	II
209	II	I	III
206	I	II	IV
202	I	III	IV
200	II	III	III
194	II	II	II
188	I	I	I

SUMMARY

Group	Two groups			Three groups			Four groups		
	No.	Mean	S.D.	No.	Mean	S.D.	No.	Mean	S. D.
I	12	215.1	11.6	8	215.1	12.7	6	214.7	13.8
II	12	215.5	10.4	8	215.1	10.6	6	215.5	11.2
III				8	215.6	10.2	6	216.0	9.5
IV							6	215.5	8.9

The summary table makes it clear that even with so small a number of cases, this method of distribution produces groups which are much alike with respect to the mean or average. How-

ever, the standard deviations (see p. 193) have been listed also, to illustrate the disadvantage of this method. The standard deviation for Group I is always higher, because Group I contains the extreme high and extreme low score. The groups nearer the middle are more homogeneous and hence have smaller standard deviations. If as small a number of subjects as this were to be used in an experiment in which the variability of the scores was important, it would be necessary to juggle the subjects around a little more nearly to equate standard deviations. An extreme score would have to be placed in one group and balanced by two or three less extreme scores in each of the other groups. Or, if more subjects were available, the elimination of the highest and lowest pupil, and the substitution of pupils with scores nearer the second and next to the last, would tend to reduce the difference. No formula can be given for this cut-and-try procedure. Sometimes it is necessary to include one pupil whose scores are markedly different from those of the rest of the group. In certain experiments it is possible to equate one pupil with two others, one higher and one lower, averaging the two and treating the average as though it were one equivalent pupil. Unless they are essential for the experiment, and unless they can be equally divided among the groups, extreme scores are best dropped entirely.

3. Use of Equated-group Methods

The equated-group method may be said to be useful because of the following factors:

(*a*) It offers opportunity to use a control group, thus making sure that one factor, and only one factor, produced the result.

(*b*) It is the only method possible for the comparison of two or more operations, each of which tends to affect, to a greater or less degree, the other. In determining whether delinquency is more affected by Sunday-school attendance or by Y M C A membership, it is necessary to have at least two groups, for the success of either in preventing delinquency would almost surely carry over into the period when the group, if only one were used, entered upon membership in the other organization. In such an experiment it would, of course, be desirable to use three

groups, in order to have one for a control group, subject to all the influences of home, city, school, etc., except the single operation of Sunday-school or Y M C A membership.

The method of equated groups has certain difficulties and disadvantages, likewise.

(*a*) It is subject to the difficulty which any limited experiment faces in the selection of groups. Because of difficulties in equating, the tendency is to take highly select groups for experiments. Then, all too frequently, the results do not correspond to those found in the wider situations in which it is hoped that results will be used.

(*b*) It is very difficult, in practical life, to equate groups so completely that small differences can be measured. If it is done with a small group, it requires careful measurements and a very precise selection from a large group of possibilities. Usually adults are too much interested in their individual affairs to enjoy the sort of restrictions imposed by experimental conditions. Such experiments are usually performed upon children, since the control over their life conditions is easier to establish. If it is sought to equate groups by studying large numbers and allowing for the chance cancellation of variables, then the groups must be very large, and must not be selected with any constant error. This is quite as difficult as the laboratory type of enterprise.

(*c*) No groups are completely equated. Certain unsuspected or unmeasured factors may be the real cause of apparent differences. Usually analysis of the change to be measured, common sense statement of expectations, and previous experience with the traits being tested, will indicate to the experimenter the things which must be equated. Thus it is usually necessary to equate for intelligence, seldom necessary to equate for length of the thumb nail. In certain experiments groups would have to have exactly the same diet, in others such a factor would be apparently irrelevant. In any experiment on teaching methods, the learning possibilities of the groups have to be equated, and also the teaching skills of the experimenters who are trying the methods. Two classes quite alike might show a spurious difference between methods, if one method were applied by a

capable teacher whom the pupils liked, and the other applied by an equally well-trained teacher whom the pupils disliked. After the best judgment has been applied, and the best equating developed, there are elusive factors which may influence the outcome. If these are present equally in both groups, very well. If they are in one group, and not the other, the results are confused.

c. Rotation Methods

This third difficulty has led to the development of a third type of investigation. This type, the "rotation method," is really trying a series of methods, in different orders, on equated groups. It is the most refined technique for experiments of the controlled, laboratory variety. It gives a fair chance for each operation to be tried out with each group.

Thus a state Y M C A secretary, anxious to find out whether Hi-Y clubs should carry on programs of Bible study or of modern-problem discussion, considered this procedure. Let each Hi-Y club be tested for religious and ethical development (T_1). Let one third of the groups try Bible study, a second third try discussion, and the final third hold regular business and social meetings but attempt no study program. At the end of 2 months let there be a retest, like the first one (T_2). Then let the operations shift. Let the first third now work at discussion, the second third have no study, and the third third start Bible study. At the end of 2 months more apply duplicate equivalent tests again (T_3). Then shift a third time, each group trying the operation it had not heretofore tried. The first third would do nothing but hold business and social meetings, the second third would embark upon Bible study, and the final third would work at the discussion of modern problems. At the end of 2 months more would come the final test, T_4.

The rotation method applied to a more complicated experiment requiring three operations and a control, may be diagrammed as follows:

Group I $T_1 \rightarrow O_1 \rightarrow T_2 \rightarrow O_2 \rightarrow T_3 \rightarrow O_3 \rightarrow T_4 \rightarrow O\text{Con} \rightarrow T_5$

Group II $T_1 \rightarrow O_2 \rightarrow T_2 \rightarrow O_3 \rightarrow T_3 \rightarrow O\text{Con} \rightarrow T_4 \rightarrow O_1 \rightarrow T_5$

Group III $T_1 \rightarrow O_3 \rightarrow T_2 \rightarrow O\text{Con} \rightarrow T_3 \rightarrow O_1 \rightarrow T_4 \rightarrow O_2 \rightarrow T_5$

Group IV $T_1 \rightarrow O\text{Con} \rightarrow T_2 \rightarrow O_1 \rightarrow T_3 \rightarrow O_2 \rightarrow T_4 \rightarrow O_3 \rightarrow T_5$

Which operation has the advantage of coming first in the list? Which operation has the advantage of any peculiarity in Group III? Which operation is most affected by a general epidemic coming between T_2 and T_3? Obviously these factors tend to be the same for every operation. Their influence is said to have been "rotated out."

Like equated group methods, rotation is practicable either on a very large scale, trusting to chance to cancel out variables between groups, or on a refined, laboratory experiment in which each item that could possibly affect the outcome has been measured and in some degree equated. Thus within one department of a Sunday school, it might be possible to select four classes, and redistribute the pupils so that the resulting four groups would be alike in average age, in proportions of each sex, in average intelligence, average home background, average knowledge of the Bible, and average years experience with the Sunday school. Then let each group be given teachers as nearly alike in ability as possible, and the same lesson materials. Let one group begin with dramatizing the stories, a second group learn the stories to tell them to others, a third group illustrate the stories by drawings and paper construction work, while the fourth group are simply told the stories with no attempt to secure activity on the part of the group. After a test for memory of the stories, let the groups shift operations, and try a second interval of study on a new basis. Here, since the carry-over from first learning to the second period might be large, it would be necessary to select new subject matter, approximately equal to that which was studied during the first quarter of the experiment. Then after the third test, let operations shift again, in the rotation fashion, until each group has tried, in turn, each operation. The differences between achievements by Operation I (dramatization) taken on the average in all groups, Operation II (story telling by the pupils) averaged for all groups, and Operation III (construction work) averaged for all, would indicate the difference in memory value between these types of expressional activity. The difference between the average for each and the average for the control group which did no expressional work, will give some idea

of the worth of pupil activity in a Sunday school like the one studied. Any slight inequalities due to practice effect, for example, would be rotated out, applying equally to each method. Any constant influences, such as the difficult disciplinary problem in one group, or the slightly superior teaching skill of one teacher, would also have a chance to apply equally to each method and hence not influence the result.

1. Use of Rotation Methods

In general rotation methods are applicable as follows:

(*a*) They are the best methods for securing really reliable results on a several-group basis, in comparing operations one of which does not much help or hinder the next.

(*b*) They tend to eliminate the error due to any variable which is constantly influencing the result. This may be in the group or the surroundings or the experimenter.

(*c*) Since rotation methods involve the repetition of several operations within each group, the carry-over from each operation to every other must be the same. Ideally it should be zero, that is, the operations should be of such a nature that having done one, the second begins back at the beginning, and the third, fourth, fifth, etc. likewise. The rotation method, however, unlike single-group methods, is still applicable when the carry-over exists, provided it is uniform. If having gone through one kind of discussion contributes as much to the second as the second would have contributed to the first had the order been reversed, then the results obtained by a rotation method, so far as comparison of the two or more operations is concerned, are reliable.

d. Survey Methods[1]

The fourth type of method to be considered is the survey method. It consists, essentially, in doing the testing while "Nature," or the ordinary course of events, performs the experiment. Thousands of experiments are going on all the time.

[1] The word "survey" as used here should not be confused with its more common meaning, a study of a particular situation or organization with a view to improvement of the efficiency with which processes are carried forward. For discussion of surveys in this sense, see SWIFT, "Survey Methods," *Religious Education*, May, 1927.

Children are growing up in homes which are quite alike, save for the fact that some are church-going, others are non-church-going homes. Which is superior? It should be possible to test a thousand children for various behavior traits, and study differences. Of course, the factor always to be watched is the possibility of difference in the groups studied. Would the groups really be alike save for their church relationships? If so, the answer is significant. Otherwise, of course, it may be that the church-going homes represent homes where different education, different books, different magazines, etc. would have produced desirable results anyhow, regardless of church. Another example may be taken from the field of organization. In some communities week-day religious education is financed and organized by each church independently. In others it is done cooperatively. What are the differences in cost, attendance, and significant results in the two situations? Again the stress should be laid on finding communities which would be alike, other things than week-day religious education considered. Is there a tendency to cheat during the freshman year at college more than during other years? Do people who have lost one or both parents in childhood tend to have greater reserve and lack of sociability? Do boys who have been given sex education prove less subject to mental conflict and delinquency than those who have not? Do people who can recite the Ten Commandments show less tendency to break them than do other persons of equal intelligence and social status? The list of questions is almost unlimited, and the persons are to be found throughout society, all ready to be tested.

1. Correlation in Causal Studies

Correlation[1] is one of the techniques by which causal investigations are conducted. If two factors are seen to be so related that whenever one is present the other is, and particularly if the two are alike in degree whenever found, then they are said to have a high positive correlation. This may be interpreted, causally, as meaning that A causes B, or that B causes A, or in most cases, that A and B are both due to the operation of some common

[1] See p. 197.

factors. Thus suppose we find, as Schwesinger has found, a high correlation between scores of children on the Orr "Good Manners Test," and on the Schwesinger test of social-ethical vocabulary.[1] That might be interpreted as meaning that knowledge of certain words brought about good manners, or that good manners brought about increase in ethical vocabulary, but probably should be interpreted as meaning that children from certain homes tend to have, because of home training, both good manners and good vocabulary.

The refinement of the correlation technique to the development of partial correlations marks a significant advance in the possibilities of experimental investigation of human behavior. A psychologist recently called it one of the greatest developments of his generation. Human beings are even more difficult to manipulate than the traditionally perverse, inanimate objects. Plants can be grown in sunlight and dark, in health and disease, under any handicaps which the ingenuity of the scientist may tempt him to employ. Not so children. A most significant study may be perfectly hopeless, if individuals (or their guardians choosing for them) do not care to undergo the scientific operations. Laboratory studies of groups of human beings are always difficult, even where there is no objection to cooperation. Variables creep in through a word read by this person, or a colony of bacteria developing in the viscera of another person. Many people have felt that the study of religious growth experimentally was so hopeless that they were forced to rely upon idle speculations. They were wrong, of course, for careful observation of the life around them was never denied them. But the development of statistical science makes possible an even more promising method of study.

By the use of partial correlation, it is possible to take individuals of varying ages, varying degrees of intelligence, varying degrees of religious preparation, varying attitudes toward the problem in hand, and predict what the relationship would have been between their Bible knowledge and their popularity in

[1] SCHWESINGER, "A Study of Socio-ethical Vocabulary," Teachers College Bureau of Publications, New York, 1926.

college, *if* they had all been of the same age, the same intelligence, the same degree of religious preparation, all had the same attitude score, etc. In short, the relationship between any two factors can be studied regarding all other variables which are measurable, as though all those other variables were constant. This amounts to equating non-equal groups statistically instead of being under the necessity of doing so in the difficult actual situation.

Thus in the above-mentioned investigation proposed by one State Y M C A Secretary, rather than equating specific Hi-Y clubs or hoping that by large numbers all variables would cancel out, the important variables can be measured and statistically equated. In truth, rather than performing a set experiment in which, to eliminate the selection factor, certain groups would arbitrarily be set at Bible and certain groups at discussion, it would be possible to let each group go its own sweet way, a course which many supervisors heartily welcome. It would be necessary only that each group keep accurate record of the time spent at Bible study and the time spent at problem discussion. Tests given at the beginning and end of the year would provide a basis for determining the gain made by each group. Then it would be possible, by partial correlation methods which will be discussed later, to see just how closely the gain was related to Bible study, and how closely it was related to discussion. This could be done in such a way that other factors like quality of leadership, intelligence of pupils, previous background, and home training would not interfere at all. They could be "held constant" or "parceled out."

2. Use of Survey Methods

Survey methods are evidently desirable in certain respects.

(*a*) They make it possible to study problems which cannot well be brought into laboratory controls.

(*b*) They make it possible to utilize the vast amount of experimenting constantly going on. The expense is a minimum because only the testing needs to be done in many cases. The operations have been performed.

(*c*) They make it possible to study factors which it might be

harmful to introduce into an experiment. Thus it would be quite unwarranted, in the interests of science to debauch a group of normal individuals, in order to study the effect upon children of alcoholic indulgence on the part of parents. The fact that many do now seem interested in becoming subjects for such an experiment, makes it possible to study the results by a survey method.

The objection to survey methods is largely in terms of the large group which must ordinarily be studied, and in terms of the difficulty of securing objective measurement of important variables. It must be remembered, however, that those same difficult measurements are presupposed in every careful experiment by whatever method.

3. General Advices

a. Miniature Experiments Advisable

Whatever the experimental method finally chosen it will be found advisable in almost every case to work through the problem on paper before the experiment is actually set under way. This involves making a good guess at the results that are likely to come from each of the measurements undertaken and going through all the statistical processes in connection with them. There is seldom an experiment in which the experimenter does not feel, after he has gotten his work well under way, that if he had only thought of it at the beginning it would have added greatly to the value of the experiment to control this or that factor or to take some simple check on the results. Oftentimes, working the experiment through on paper is not sufficient. It is necessary to try it out with a small group. Any experiment which is to be carried on on a large scale should certainly be tried out first on a small group, everything being done just as it will be done eventually in the large group, every result being figured out with great care until the final conclusion is reached in precisely the form in which it is hoped to be reached on the basis of the large group result.

b. The Diary

One of the most valuable additions to an experiment is the keeping of a careful day-by-day log of the enterprise. The diary

may well begin with the early formulations of the problem, and the observations which preceded any carefully controlled experiment. Any distractions which entered during testing should be noted. Weather conditions, attendance reports, the examiner's headache, bad news received by one of the subjects, or the excitement of a holiday season: each of these may shed some light on what happens.

From time to time slight changes may creep into the procedure. These should be recorded wherever it is conceivable that they might influence outcomes. The experimenter, noticing these details, is inclined either to rule them out as insignificant, or to believe that he will remember them and make the necessary allowance in his report. Either is a dangerous assumption. Memory is subject to many slips throughout a long experiment, and the inclination of the experimenter at the moment is not a safe guide to what other people would later regard as important and relevant. Perhaps it is unnecessary to mention the experience of one research worker who found that he kept his diary in so much haste that many of the notes, and even certain dates, could not be made out, when he came to write his report.

c. Experiments with Individuals

All of the suggested methods have been illustrated by application to group experiments. It is worth noting, however, that the techniques of experimenting with individuals are not essentially different. The dean of men in the university may take one man as his "group." One operation or several may be tried. With several individuals equated, various forms of discipline may be compared and even rotated.

"Survey" methods may be employed to study how certain experiences within the individual are related to other experiences. A girl had had four nervous breakdowns. Asked to discover a factor which had been present just before each breakdown, she found that each had come when her interest in some married man began to arouse comment. It required no further experimental operations to make clear to her and to her counsellor that a causal relationship probably existed.

Because the objection is sometimes made that experimental

and statistical techniques fit the crowd but not the individual, and the conclusion is therefore drawn that they are not useful for those whose supreme concern is the welfare of the individual, something may well be said on the applicability of experimental results. Of course, experiments made upon individuals may be applied to those individuals. The difference between a careful scientifically measured experiment and the hit-and-miss experiment of the well-wisher, is comparable to the difference between the diagnosis made by a neurological institute after thorough X-ray and chemical examination, and that which would be made by a well-meaning quack in the field of health. The contribution of experimentation to individual welfare is, however, wider than that of individual diagnosis and prescription. An individual is a person because of social experiences. Insofar as those experiences are controlled by religious workers, they are usually group controls. Sermons, classes, club meetings, discussions, articles, stories, dramatizations, plays, recreational facilities, games, moving-picture shows, newspapers, and financial campaigns influence whole groups at once. At almost every point we find that the effort of the social-religious worker is directed immediately toward groups; not groups, as ends in themselves perhaps, but groups as the real setting through which individuals are reached. To find out truths about groups is to enable the servant of humanity interested in individuals to serve individuals better through the many channels by which he is actually influencing individuals.

EXERCISES

1. Select a problem for experimental investigation during this course. Justify it in the light of the principles suggested in this chapter.

2. Indicate the objection to each of the choices listed below, by placing in the parentheses following each, the number of the principle which is violated in the choice.

a. What are the causes of war? ..... ( ).
b. Do intelligence tests predict school success? ..... ( ).
c. What words in the book of Haggai are outside the vocabulary of ten-year-old children? ..... ( ).
d. An investigator interested in adolescent girls is persuaded to try out a new technique for determining the relation-

ship between school success and success in the ministry, in order to secure credit toward a degree..... ( ).

e. The XYZ Bible School starts an investigation to prove the value of the Bible in the life of the individual..... ( ).

3. Consider the list of suggested problems in the Appendix. Choose one which could be answered by the single-group method, and indicate how the experiment would be carried on. Assume the existence of adequate tests. Find likewise an example of the equated group, rotation, and survey method from among those suggested.

4. Suggest an important problem for religious education not included in the list of 140 problems, and indicate the type of method used for solving it.

5. Suppose it is desired to find out the comparative value of long and short hymns (40 *vs.* 12 lines) in connection with adult services of worship. How would this be done with a single group? With equated groups? By rotation? Could survey methods throw light upon it?

6. Fill in the blanks in this statement. A correlation of _____ between length of hymn and enjoyment of hymn would indicate that long hymns were enjoyed _____ _____ than short ones.

7. What type of method obviously could not be used to investigate each of the following elements? Refer to the limitations upon each method suggested in the chapter.

a. Does a given family enjoy auto rides more than movies?

b. Is a motto more effective in producing honesty in school work than would be a story?

c. Do Italians need a different sort of worship from that needed by Anglo-Saxons?

8. Consider the following activities of a religious organization. Underline each one in which knowledge of the average tendency of groups would be more useful than knowledge of each individual, without any combination into group results.

a. Choice of best style of print for church bulletin.

b. Selection of a week for special meetings.

c. Choice of one illustration rather than another for a sermon.

d. Choice of best form of address for a given personal letter.

e. Choice of Christmas presents to be given out in Sunday school.

f. Selection of vocabulary for a study book.

g. Choice of officers for a particular committee.

CHAPTER III

METHODS OF MEASUREMENT

1. Observation as Measurement.

a. Whatever Exists Can Be Measured

A story is told of a group of schoolmen who, in the early days of the measurement emphasis in public education, had come together to talk about tests. One superintendent was speaking fervently his opposition to the notion that the important aspects of mental life could be measured by any man's footrule. He closed his protest with the impassioned question, "Who would presume to measure the intellect of a Milton or a Shakespeare?"

It is said that Thorndike arose to answer him. "Fortunately it is not necessary for us to measure the intellect of a Milton or a Shakespeare. That has already been done by the previous speaker. They have not only been measured, but placed at the head of the list. It now remains only to find out where the rest of the human race stand with reference to them."

It may be vain to hope that measurement in religious education will be spared similar attacks by those who do not realize that it is not the measurement, but only the refinement, which is a recent development. Measurement is older than language, and contrary to certain assumptions, it was probably the intangible personality traits which were among the first to be measured. As soon as there arose a feeling that I am abler than he, that this mate is preferable to that, or that this food, drink, sunshine, or sport is more satisfying than certain other foods, drinks, states of warmth or boredom, then there was a rudimentary measurement. "More" and "less," "better," "bigger," "stronger," "purer," all imply amount, and degrees of amount. If it can be said of any attribute of religious life that A possesses more than B, or that C has attained a more desirable state than D, then measurement is already in operation. From this point of view it

becomes difficult to understand just what is meant by people who talk about the importance of the things which cannot be measured.

It is true, indeed, that many of the important aspects of life have not been measured by any very refined instruments, as yet. It is probably true that the more refined techniques will not be applied satisfactorily to some of these traits for many years. It is perhaps well to refrain from hasty construction of measuring instruments and from their widespread application at present. Yet these limitations must be viewed with regret by any who would not hold a brief for crudity, error, and unreliability. Measure we almost surely will. The question raised by the experimental approach to religious education is, "Can we not measure more accurately, carefully, and reliably so that results will be more dependable?"

b. Methods of Refining Crude Observations

Let us suppose that an experimental problem involves knowledge of the effect of certain methods for producing honesty in children. If the methods are to be compared, it will be necessary to measure the degree of honesty in certain individuals or groups before and after the methods are applied. It has already appeared that this can be done, and perhaps most frequently is done, by the judgment of some person that the children have improved or failed to improve. There are many steps which can be taken toward making this judgment or rating more useful.

1. Judgments may be improved by increasing the expertness of the judge. To evaluate honesty is no mean task. Other things being equal, a judge who has had much experience in evaluating honesty and checking up on his ratings is preferable. A rater who has participated with others in discussing just what honesty is and how it is indicated will probably be more expert than one who has made no such study.

One of the most important aids in training judges is to dispel unfounded notions of symptoms. Raters should be aware that there is no basis in fact for the expectation that large noses, prominent chins, or bulging foreheads reveal character. Holling-

worth[1] submits sufficient convincing evidence, but other investigations later have added to the accumulation. Professional "character readers" have been tried and found wanting. Katherine Blackford's traits of "blonde" people, which supposedly result in good salesmanship, have been shown to be equally often present in brunettes.[2] Other careful experiments have dealt with the ability of competent persons to estimate intelligence from photographs. Students, professors, employment managers, professional character readers, all were given ample opportunity to arrange in order of intelligence the persons about whom they had no evidence except the photographs. The results showed that some groups would have done slightly better if blindfolded! Mere chance agreement and disagreement, nothing more. It has been well said that the bumps on a man's head tell more about his wife's character than his own.

2. Judgments may be improved by increasing the number and variety of contacts which the judge has with the subjects being rated. Experiments have indicated that two school teachers rating a boy will agree more closely than will a school teacher and a playground instructor. Teachers of the same subject rate more nearly alike than do teachers of different subjects. In other words, each pupil has a form of behavior for each social situation. Any judge who sees the pupil only in Sunday school, only in arithmetic class, only in his home, only on the playground, or only when reading in the library, has an inadequate picture. Moreover, the same individual in the same general setting, will respond differently at different times. The judge who has been acquainted with the pupil for 3 years may have a basis for judgment which is not possessed by the judge whose experience is limited to 3 hours' observation. The number of significant contacts is also increased by the knowledge of the judge that he is expected to give a rating. It may well be that a judge who was preparing definitely to rate a child on honesty might render

[1] HOLLINGWORTH, H. L., "Judging Human Character," D. Appleton and Company, New York, 1923.

[2] WINTER, "Blonde and Brunette Characteristics," *Psychological Bulletin*, p. 655, November, 1925.

a better judgment at the end of a week of observation than would be rendered by an equally able judge who had lived with the child off and on for 2 years, but with no intention of observing any particular trait.

Knight[1] has pointed out an interesting exception to the general principle that ability improves with acquaintance. Long and intimate friendships tend to bring marked decreases in the trustworthiness of ratings. Close friends are overrated on desirable traits and underrated on less desirable traits.

3. Judgments may be improved by multiplying the number of judges. Probably it is better to take the median or middle rating of the series, rather than to average all marks. There is a limitation to this principle, in terms of the above criteria. One expert judge with many contacts may be preferable to three casual and inexperienced raters. However, Rugg[2] is very certain as a result of his army experience that unless the judgment of at least three competent persons is secured, the rating will be very undependable.

4. Judgments given by persons who are entirely disinterested are preferable to those given by persons whose prestige is involved in the experiment. This seems almost too obvious for mention, yet observation of reports in religious papers, of experiments tried by churches, Sunday schools, clubs, and other groups shows that two-thirds of the reports ignore this item. The success or failure of the activity is stated in terms of the opinion of the person who thought it up and put it in operation. Such records are of little value for the scientific appraisal of experiments.

5. Judgments may be improved by breaking up a general quality into its component parts. Thus Furfey[3] found that when judges were asked to rate the social development of boys, the reliability[4] was only 0.75, but when that general "social

[1] KNIGHT, "Effect of the Acquaintance Factor upon Personal Judgments," *Jour. Ed. Psy.*, Vol. XIV, pp. 129–142.

[2] See discussion on pp. 45–6.

[3] FURFEY, "An Improved Rating Scale Technique," *Jour. Ed. Psy.*, p. 45, January, 1926.

[4] See p. 44.

development" was broken up into 18 units each of which was rated, the reliability of the total rating went up to 0.90.

There is a limit beyond which this analysis of a general trait into specifics is not useful. Hartshorne and May have found that ratings on general honesty are better indications of tendency to cheat than are ratings from the same teachers on the specific tendency to cheat. Rugg and Slawson have found that "general all-round value" of officers and teachers can be more reliably rated than can some of the specific things which make it up.

In the Indiana Survey of Religious Education,[1] there is proposed a score card to help judges rate curriculum materials. General worth has been separated into a large number of specific items. For each item a scale has been made, as described above. The question has been raised, but so far not answered, "Does analysis so elaborate as this, really increase the reliability of rating, beyond the simple estimate of worth?" Opinions differ, but are quite useless until experimental evidence is secured.

One of the principal advantages of breaking the general up into specific parts is that each part may be given its proper weight, and hence the emphasis of different judges corrected. In judging a worship service one judge may be unduly depressed by the leader's necktie, another unduly happy over a favorite hymn. A score card would break up the lump judgment into such items as "atmosphere of room," "leader's appearance," "leader's voice," "material chosen," "suitability of theme to pupils," "suitability of hymns to pupils," "suitability of hymns to theme," etc. The number of points which could be given each item would be limited. No single item could thus be emphasized to the neglect of other more important items.

Another illustration of the score card may be found in the instrument for evaluating city church plants, used in the Indiana Survey, Volume II. One thousand points are distributed by careful judgments to give the proper weight to every item from worship room to toilet facilities. This particular score card, however, along with many another, is subject to one serious error.

[1] "The Indiana Survey of Religious Education," George H. Doran Co., New York, Vol. II, pp. 101–337, 1924.

It is assumed that, since 200 points may be awarded for religious schoolrooms, and only 20 for gymnasium, the schoolrooms have been given ten times as much credit and influence on total score, as the gymnasium. This is not necessarily true. If in marking church buildings, they range from 180 to 190 on schoolrooms and from 0 to 20 on the gymnasium the second will probably be twice as influential in determining total score as is the first. It is not absolute size, but relative size that matters. Speaking statistically, for those who like it that way, it is not the size of the mean which is influential, but the size of the standard deviation. It is the variability or spread that weights an element which is to be added to others in a grand total.

This difficulty may be remedied in the following manner. Suppose that the general trait, whether of person or building, be divided into its various parts. Suppose that each part be rated on a line graph, like this.

Place an x on the line to indicate the position of . . . . . .

Poorest imaginable	Very poor	Poor	Fair	Good	Very good	Best imaginable

Then the position of x may be given a numerical value in terms of a scale which is laid along the line. In case the element is very important, the value may be multiplied by a weight of 2, or 5, or 7, or 10, or 20, to place it in its proper proportion to other elements which have been rated on a similar graphic scale. Weighting by multiplication takes care of both absolute and relative differences.

6. Judgments may be improved by finding specific, objectively verifiable questions which give evidence on the matter being judged. Thus rather than ask whether or not a room is light, it is possible to ask the window space per person or per cubic yard of room. Rather than ask whether Johnny is neat or not, it is possible to ask, "Does he get ink stains on his hands?" and "Does he keep his books in order in his desk?" and "Does he keep his room straightened up?" While it would be unfortunate if the essential spirit of a characteristic were omitted because it could not seem to be stated objectively, yet it is surprising how

many apparently intangible qualities really do have specific signs or tokens which can be sought. The Upton-Chassell Citizenship Scale[1] furnishes many excellent examples of character traits stated in terms of objective manifestations. The Colgate Mental Hygiene Test[2] and various health habit charts make further suggestions.

7. Raters should be aware of the tendency which Thorndike has called the "halo" effect. This name is given to the human characteristic of reacting in toto, rather than to analyzed parts. One has a general set or liking toward Dorothy. Therefore it is easy to believe that Dorothy is honest, brave, kind, sweet, beautiful, and healthy. One dislikes Jeanette and therefore readily suspects the worst and overlooks many fine points. One investigator[3] reported a correlation, when raters had been judging prospective teachers, of 0.95 between moral character and quality of voice. The relationship was undoubtedly a matter of the "halo" in the mind of the rater. Ratings should be examined to see the extent to which this factor has entered. In general the best ratings are indicated by a high correlation among various judges rating the same trait with a very low, but probably positive, correlation between the ratings given on one desirable trait and the ratings given on every other different, but desirable trait.

8. Judgments are sometimes faulty because the ratings are bunched at one end of the scale. Thus deportment grades in school are usually too largely "Excellent" and "Good" to make them useful as possible character indicators. There is a tendency in applying a scale to human character traits to give a skewed distribution, with a large group near the top and only a few scattering to the lowest extremity. Certain devices have been suggested to correct this.[4] Perhaps the best suggestion is that a graphic rating scale be used, similar to the one suggested above, in which a check mark is placed along a line. It is suggested that

[1] See p. 95. [2] See p. 75.

[3] KNIGHT AND FRANZEN, "Pitfalls in Rating Schemes," *Jour. Ed. Psy.*, Vol. XIII, p. 204, 1922.

[4] For samples of improved rating scales, see WATSON, "Rating Scales," Occasional Studies No. 2, 1927, Association Press.

if possible on a single sheet, the entire group be rated on this one trait, rather than trying to rate all the traits in one person, on the same card. Moreover, Symonds[1] suggests that over each division of the scale there be indicated the proportion who might be expected to fall within that general section. Thus a scale for initiative might consist of a sheet headed as follows, with the names running down the sheet, and a line after each name.

TABLE II. — SUGGESTED FORM FOR GRAPHIC RATING SCALE FOR INITIATIVE

DIRECTIONS: Rate each pupil by placing a cross on the line at an appropriate point. Try to keep the proportion of crosses within each column fairly near the per cent suggested.

	Purely a follower. Helpless	Seldom starts anything new by himself	Usually follows others. Does take some lead	Ordinary occasional initiative	Apt to be good leader	Original, good ability to get things going	Always in the lead Never at a loss Very original
NAME	About 2 per cent	About 6 per cent	About 20 per cent	About 44 per cent	About 20 per cent	About 6 per cent	About 2 per cent
Aden, Frances							
Ball, Billy							
etc.							

9. Judgments become more useful when it is known just how certain the rater is of his rating. Studies have demonstrated that ratings of which the judge expresses himself as "very certain" are apt to have a reliability of 0.80 to 0.90 as contrasted with those of which he expresses himself as somewhat doubtful which may lie in the range between 0.20 and 0.60. It is not necessary that the judge be able to give a clear and definite reason for the rating. Landis[2] has shown that ratings are sometimes very good indeed, although the judge is unable to state any good basis

[1] SYMONDS, "Notes on Rating," *Jour. App. Psy.*, 1925.

[2] LANDIS, "Study of Ratings," *Jour. Personnel Research*, Vol. IV, p. 6, May 1925.

for the faith that is in him. He found little or no correlation between ability to give good reasons, and ability to make good ratings. There is, however, a very high correlation between reliability of ratings and the emotional conviction of the rater that his rating is valid.

10. Judgments may be much improved by clear definition of the trait and the use of samples to illustrate degrees. Thus in the illustration, honesty might well be defined to include certain ranges of activity. If it were desired to include under such a term the willingness of a person to recognize honestly his own shortcomings, or his ability to detect the dishonest character of investment in corporations whose practices are legal but questionable, then it should be made clear that such interpretations are wanted. Many judges might not seek honesty in those particular realms unless so directed. It should go without saying that the definitions should not be more complex than the thing defined.

The sample scale has been developed in handwriting, drawing, and literary composition, to improve ratings by illustrating various degrees of excellence. Hundreds of samples of compositions, handwriting, specimens or drawings, were graded by hundreds of judges. As a result, the specimens could be arranged in order of excellence, from poorest to best. It was found possible to go even further, and find the distance between each two samples, numerically. Then specimens were printed. Any new sample might be graded by holding it alongside the scale until its position with reference to the other samples was clear. It might be found better than sample rated at 9.31 but not quite so good as 9.45. Rough interpolation could give a very fair rating to this sample. It was found that several judges agreed with themselves far better if guided by such a sample scale than they would if left to their own devices to evaluate the new specimen.[1]

The same principle is applicable in the measurement of other qualities. Sample "honesty" situations may be set up, and evaluated by several hundred judges. From these a scale may be built up, upon which future comparison can be made. Thus

[1] AYRES, "A Scale for Measuring the Quality of Handwriting of School Children," No. 113, Russell Sage Foundation, New York City.

the judge working with such a scale could say not merely, "I think he is more honest than he was a year ago," but could say, "His actions a year ago were capable of being rated at about 6.27, whereas his acts now are better than those marked 9.68 on the judgment scale." A detailed description of the method used in constructing such a sample scale and assigning values to the samples, is given in the Appendix.

A simple application of the idea of using samples to guide ratings was worked out by Scott[1] in the army rating scale, sometimes called the "man-to-man scale," or the "human-ladder" scale. For each trait, the raters were asked to choose a man who represented the best individual they knew. Another was suggested to go at the foot of the scale. The highest man was given a value of 15, the lowest a value of 3. Halfway between those two illustrations stood the ordinary, average, common, or garden variety of individual. A third sample was suggested to represent this center of the scale, and assigned a value of 9. Similarly each rater chose one individual who symbolized for him the points 6 and 12. Then each of the men to be rated could be compared with the samples who had been located as symbolizing the fixed points. A new man, slightly better than the fellow who was rated 6, but decidedly below the fellow who stood for 9, would be rated 7, and so on.

c. Value of Ratings

Suppose these precautions all to have been taken? How valuable are ratings obtained under such circumstances? Unfortunately this question cannot be satisfactorily answered at present. In the first place there are few investigations which have been careful to use trained judges sufficient in number, disinterested, rating on a trait which had been analyzed and properly weighted, utilizing objective indications wherever possible, correcting consciously for the halo effect, yielding ratings properly distributed over the entire range of the scale, expressed in terms of degree of certainty, and compared with suggested samples. Knowledge of what 'twere good to do has far outrun practice in this sphere.

[1] Scott, "The Army Rating Scale," *Psy. Bull.*, Vol. XV, p. 283, 1918.

In the second place it is difficult to find good tests of the efficacy of scales. Some would check them against standardized tests, others would hold that ratings built up in accord with the 10 principles suggested are a good criterion against which to check tests. In case of disagreement, shall we believe the tests or the judgments?

d. Reliability of Measures

One standard always applied to scales and tests is that of "reliability." Reliability answers the question, "If the test were given again under the same circumstances, would it yield the same results?" Frequently also, it is the answer to the question, "If the ratings were made by other judges, would they agree with the first judges?" In its first form, it is like asking, "If a person is weighed, and then immediately gets weighed again, nothing having changed his real weight in the meantime, will he weigh exactly the same?" The reliability for a government scale would be high. For a spring balance it would probably be fairly low. In its second form, reliability may be illustrated by asking "If this person were weighed by several different persons, on several different scales all purporting to measure the same thing, how closely would the results agree?" Anyone may obtain an answer by walking along a city street, investing a penny in every weighing machine that offers itself. Or he may measure the reliability of thermometers by going to a store where a large number are hung together, awaiting sale. The variations are usually sufficient to convince anyone that all unreliability is not confined to mental measures.

Reliability is usually expressed in terms of self-correlation.[1] Correlation means agreement. An agreement of 1.00 would be absolutely perfect. An agreement of 0.00 would be pure chance, high ones in the first list rating sometimes high and sometimes low in the second list. An agreement of -1.00 would mean that everyone high in the first score list was equally low in the second, and *vice versa*. It is perfect agreement turned upside down and inside out. Usually the correlations for reliability of ratings and tests are better than chance. There is some tendency, at

[1] See p. 195.

least, for persons who stood high in the list the first time to stand high the second, and for persons who stood low the first time to stand low the second time. So the reliability of an ordinary school examination is apt to be about 0.50. The reliability of ordinary standard tests is apt to be about 0.80. The reliability of excellent tests and unusually long tests, may run above 0.95. Table XXXII suggests a good method of interpreting reliability figures. The per cent given in the second column indicates the per cent of factors which were common the first time and the second time. Thus if the reliability is found to be 0.72 it may be concluded that about 51 per cent of the factors which produced the first score were also present and operating in the production of the second score. The remaining 49 per cent varied from time to time and test to test, dependent on minor shifts in attitude, health, ventilation, eyesight, or other uncontrolled variables.

Evidence upon reliability of character ratings is very conflicting. In many ways the most thorough-going study of ratings was made by Rugg,[1] investigating the usefulness of the Scott man-to-man scale for army officers. He had 100 picked men of average intelligence "*B*," who had lived, slept, eaten, played, and drilled together for 11 months. They were carefully trained during a 3-day conference in the technique of rating. Each man built up his own scale on the basis described above for constructing such a human ladder. Then each man was asked to name those persons within this group of 100 whom he felt competent to rate. Also each man was asked to name those persons whom he would regard as competent to rate him. Only double competents, persons who believed themselves competent and who were believed competent, were finally selected. Yet with all of this precaution, it was found that the ratings were discouragingly unreliable. In choosing sample men the same man was sometimes chosen as a perfect example of 15, 12, 9, 6, and even 3. On a scale of 80 points range, one man would commonly vary from another rating the same individual by as much as 30 points. On an ordinary scale of 100, a rating of 75 could be

[1] Rugg, "Is the Rating of Human Character Possible?" *Jour. Ed. Psy.*, Vols. 12 and 13, November, 1921, to February, 1922.

interpreted as meaning that one could be practically certain that the man's true value, if rated by a very large number of judges, would fall between 51 and 99, not a very specific meaning. In the case of ratings on intelligence, these ratings could be checked against actual test records. Rugg found that 10 per cent of those placed by ratings in the lowest group, really belonged according to the army tests, in the highest, or next to highest, of the five groups. From such data as these Rugg was led to conclude that the method of measuring human character by ratings might well be abandoned in favor of attempts to construct tests. Said he, "A single rating by a single school officer will only rarely place a pupil in his proper fifth of the entire group." More recent evidence is much more encouraging. Barr[1] reports reliabilities of ratings ranging from 0.40 to 0.80; Freyd[2] reports them from 0.52 to 0.87; Webb[3] from 0.56 to 0.81; Knight and Cleeton[4] from 0.80 to 0.90, Shen[5] from 0.62 to 0.91, and Furfey[6] from 0.70 to 0.97. Many of these are higher than are the reliabilities of the ordinary group tests. Of nine experimenters reporting during 1926 the use of ratings of character traits, five of them secured reliabilities over 0.60.[7] It may be that with improved technique, ratings will be found more useful for the present generation than are tests.

Several precautions need to be borne in mind when comparing ratings. One is that anybody can have reliabilities as high as he pleases by using a large number of judges, and discarding the ratings of all who disagree. A second is that reliability is related to range. Suppose two judges were asked to rank in order of

[1] Barr, "A Study in Social Rating," *School and Home Education*, p. 406, Vol. 41, 1921.

[2] Freyd, "The Graphic Rating Scale," *Jour. Ed. Psy.*, pp. 83–101, Vol. XIV.

[3] Webb, E., "Character and Intelligence," *British Jour. Psychol.* Mono. Supp., pp. 1–99, 1915.

[4] Knight and Cleeton, "Validity of Character Judgments Based on External Criteria," *Jour. App. Psy.*, p. 215–231, Vol. VIII, June, 1924.

[5] Shen, "Validity of Self-estimates," *Jour. Ed. Psy.*, February, 1925.

[6] Furfey, "An Improved Rating Scale Technique," *Jour. Ed. Psy.*, p. 45, January, 1926.

[7] Watson, "Character Tests of 1926," *Vocational Guidance*, pp. 290–309, April, 1927.

intelligence a group of 10 fifth-grade pupils, all of whom were doing *B* school work. Suppose then that the same two judges were asked to rate the intelligence of 10 children the lowest of whom was an idiot, the highest a supergenius, with the other 8 scattered at fairly equal intervals between. It is clear that the agreement would be much closer in the second situation. Great similarity in the group being rated makes it very difficult to get high reliability. Great differences in the group make it easy to get high agreement. One investigator reported spuriously high reliabilities because he used only extreme groups — the very high and the very low, eliminating the middle section entirely.

e. Validity of Measures

Reliability answers the question of the agreement of a test with itself, or the agreement of ratings by one judge with ratings made a little later, or by a similar judge. There is another question, equally important. That is, "What do the tests or ratings really measure?" One thermometer may agree with another very well, but both of them be wrong. In any case it would be unfortunate to regard the results as measures of air pressure. The answer to this question of the real interpretation to be placed on the result, is called the *validity* of the measurement. It may be that a test which looks like a test of ethical judgment is really a test of vocabulary. It may be that a test which is supposed to measure religious attitudes is really a measure of conformity to adult expectations. It may be that ratings which are supposed to be measures of initiative are really in large degree measures of physical health and intelligence. It may be that ratings on a scale for "reverence," are really measures of the extent to which the pupil does not cause the teacher trouble.

1. Methods of Validation

There is no task in the field of measurement which is more in need of study than that of the establishment of validity. Four types of validation have been commonly used. Progress is needed both in the invention of better criteria and in the refinement of the use of these existing forms.

(*a*) Measurements may be validated by checking them against other measurements, supposedly better or more widely accepted. It has been customary to validate new intelligence tests by checking them against the Stanford-Binet or the National Intelligence Tests. Not infrequently attitude tests are validated by checking them against self-ratings or the ratings of other persons. The tests of ethical knowledge developed by the Character Education Inquiry were validated mainly by checking each form against a large battery of tests of the same general sort, the assumption being that the total result of such a battery would tell more about a person than any group of judges could know in advance. Indeed, it is becoming less and less acceptable to check the validity of tests against judgment criteria. The writer, in attempts to validate tests of prejudice, found that after judgments had been made by intimate friends of the subject, if those friends were shown the test result they would abandon their original judgment, saying, "Oh well, I didn't suppose he would say a thing like that. If he did, of course, I was mistaken about him." It is astonishing to note in the literature on character tests the instances in which tests have been carefully constructed, and then checked against the loosest sort of ratings for criteria. In many cases, and the writer is not guiltless, not one of the 10 established methods of improving ratings has been utilized.

Sometimes conditions make it possible for a very intimate study to be made of certain individuals who are tested. The agreement of the test with exhaustive judgments about and experiences with a limited group is often sufficient basis for its acceptance as more widely valid. Tests and ratings are usually combined in such studies.

(*b*) Objective criteria are occasionally available for the validation of ratings and tests. Suppose it is desired to make tests which will predict academic success, ability to typewrite, or to make money. It is possible to try out the tests and see whether or not the individuals rating high do tend to succeed in school, typewriting, or money making. Probably the fact that a person has been three times in prison for arson and is again convicted of setting fires is sufficient evidence that he may be regarded as

classified with reference to a certain trait. An individual who has a steady record of having been put out of basketball games for personal fouls, is probably a good individual upon whom to study a test of good sportsmanship. In the emotional realm physiological reflexes, heart beat, blood pressure, breathing rate, and electrical conductivity may be used to check up less objective measurements.

(*c*) Sometimes validation has proceeded upon the comparison of groups. Thus Lentz studied various character tests to see which ones would show real differences between a group of boys in a reform school and a group of equal intelligence, equal age, and equal home background in a public school of the same district. It was assumed that the valid tests would be those which most clearly differentiated between these groups. The writer appraised prejudice tests on such assumptions as that a valid test would show Roman Catholics prejudiced in favor of Roman Catholicism rather than Protestantism, and that the reverse would be true of Protestants. A number of judges were used to see how widely it was believed that older ministers were more apt to be conservatively prejudiced than younger ones, that theological students at Union Theological Seminary were more apt to be prejudiced in favor of modernism and less apt to be prejudiced for fundamentalism, etc., than were Methodist ministers in a midwestern state. Some of the best validation on this basis has been done with occupational tests. If a group of the most skilful persons in the plant be selected, then a group of average workmen, and another group of workmen so poor they have had to be discharged, a good vocational test should show clear differences between these groups. An "intelligence" test which showed no marked differences between college professors of mathematics and pupils who failed the fifth grade would be rather absurd. The difficulty is, of course, that most groups overlap, and that the method is very crude. It is seldom possible to differentiate by consensus of judgment more than three or four groups at distinct points along a scale from poorest to best. It is not always safe to assume that ratings or tests which do separate these extremes, are valid for the finer, intermediate

distinctions. Moreover, if groups are widely separated on one trait, they may be widely separated on many others also. One group may be honest, the other dishonest. Yet the test which shows one group high, and the other group low, may be a test of intelligence, or home background, or aggressiveness, or reading ability, or some other quality on which the groups differ quite as truly as on honesty.

(*d*) A fourth method of studying validity is to eliminate from consideration certain possibilities. A test which appears to be a test of interests may be in reality a test in vocabulary or intelligence. Standard intelligence tests and vocabulary tests should be given to the same pupils who have taken the interest test and it should be found to what extent the test score is really dependent upon intelligence or vocabulary. In tests which are rather long and which involve complicated directions, validity should be checked by giving to some groups, tests for speed and comprehension of reading, and ability to follow directions. When these are held constant, the difference that remains may well be attributed to the sort of thing the test is striving to measure.

(*e*) A fifth method of validation, infrequently used, is that of observing change under stimulus. Too little is known about the exact results of stimuli to make this method certain. Again referring to the studies in prejudice, it was felt that the fact that education and discussion with reference to a question tended to decrease score was evidence that the tests measured prejudice rather than enlightened consideration. It is probable that a test of ethical vocabulary which could show no difference between pupils before and after the careful learning of 100 new words, would be invalid. There might be a good case made even here, however, for the notion that the test was right and the teaching method defective. Ratings "before" and "after" usually give far too much credit to changes which are supposed to have taken place.

Perhaps in this connection it is well to mention the need for long-time records of test scores. The makers of tests in religious education are so new at the business that they have not had

opportunity to follow groups of subjects for periods of 5, 10, and 20 years, to observe the changes which have taken place in test scores due to the operation of certain environmental factors. There is much justification for a sober judgment that nothing would do more to further the progress of religious education today than a system of records which would make constant report of the activities and interests, successes and failures of individuals over a period of 20 years. The record systems in most church schools are useless from this point of view. If it should prove that persons who made high scores on a certain kind of test tend to become citizens of high social usefulness, whereas pupils who made low scores tend to be arrested for bank robberies and bootlegging, that test would have an established validity with reference to conduct, which no test at the present time can have. The first step in prediction is record taking. There must be one or two generations of persons who are willing to keep records before we will have a generation of scientists in the realm of religious education who will be able to predict and control forces which the world is eager to use today.

2. Validation May Be Unnecessary

There is another viewpoint with reference to validity which deserves serious consideration. Its exponents would say that it matters little what we call a measurement. The real question is what that measurement will do. If it is found to be highly reliable, it can be used. We recognize that the study of electricity is possible and fruitful purely in terms of what it will do to certain instruments. It is not necessary to capture, analyze, and define the electric current before it is measured. They would say that the question of whether or not the Stanford-Binet measures what somebody means by his definition of intelligence is irrelevant. The Stanford-Binet measures something very reliably. We have found out many things about the relationship of that something to other abilities. We have found that it tends to improve at a certain rate, at least up to 16 years of age, the rate being fairly constant. We have found out just how well it will predict school marks. We have found out that people who make high scores are more likely to succeed in

business than are some other people. Call this intelligence, or alpha, or "Binetism" or what you will. It is something about a person which can be reliably measured and which it is useful to know.

So it might be said of character tests that endeavor to match up our English vocabulary with actual psychological reactions is vain. Who knows that there is any unity in such traits as "trustworthiness," "suggestibility," or "obedience"? Indeed it seems very probable that these traits vary from one situation to another within the same individual. A man may be a different person in his business office from the person he is when playing with his children at home. Voelker found the trustworthiness with which a boy responded in one situation to be a very poor index of what would happen in the next. After a boy had been measured by 10 of Voelker's life-situation tests, were he then to be measured by a second 10, under similar circumstances, the prediction of his second score from knowledge of his first, would be only about 5 per cent better than sheer chance.[1] In the case of ratings it may matter little whether the teacher's rating was influenced too much by the shape of pupils' noses or the tone of their voices, so long as that rating is dependable, and can be shown to predict with reasonable accuracy how well that pupil will get along as a salesman. It is important to know that ratings by faculty members predict teaching success better than do intelligence tests, regardless of whether those ratings were genuine measures of "initiative," "poise," "judgment," etc. It is valuable to know that money success after college, ability to make marks, social fitness of defectives, school popularity and similar characteristics have been predicted more accurately by rating scales than by tests. It may be the very fact that ratings are somewhat invalid, are somewhat colored by halo effects and the like, that makes them useful in predicting social achievements and attitudes. It is quite possible that the

[1] Applying the method of odd-and-even splitting of the test to the tables published by Voelker ("The Function of Ideals in Social Education," Bureau of Publications, Teachers College, Columbia University, New York), the writer found a self-correlation of 0.30 (see p. 161 and, for interpretation, Table XXX on p. 275).

"errors" from a strictly scientific point of view, in a general estimate by another person as contrasted with careful objective tests, are really very significant. If all through life, every contact of that person with others is likely to be influenced by the same "errors," then the error may be just the thing that should be taken into account, and included in the original measurement. We often appreciate better than we analyze.

2. Objective Records as Measures

We turn now from ratings toward those more objective measures which have come into prominence in recent years. It is probable that the development of the science of religious education and character formation will require more refined instruments than judgments and ratings at their best. Three additional forms of measurement suggest themselves as significant.

The first is based upon recorded facts which have some significance. The days a person has attended school during the year; the minutes he will work at an assigned lesson without being watched by an adult; the pounds of weight he has gained or lost; the rate of heart beat; the reaction of his blood to certain chemicals; the outside activities in which he participates at school; the number of older brothers or sisters; the number of engagements, marriages, and divorces; the number of arrests; the nicotine stain on his fingers; the income he can earn; the number of persons he counts as friends; the number of jobs he has held within the past 2 years; the hours per week spent at study or movies — these and a vast number of other pieces of information constitute fairly objective measures of certain attributes of his personality. They are objective because other investigators seeking the same end, would reach very nearly the same result in measuring them. They are valid, in so far as they are not interpreted beyond their real significance. Often, they constitute a better index of the success of an experiment than would ratings or tests. May found that hours per week spent in study predicted academic success better than did intelligence tests.

The national staff of the Y M C A was interested in comparing two administrative methods. In such a situation there were

several fact ratios which might serve as tests. For example, the sections, one of which used a county unit, the other of which used a larger state division, might be compared on the number of boys enrolled. They might be compared on the ratio of the number of boys enrolled to the total Protestant population of the area, the ratio being more significant than the original figure. So with finance, the ratio of contributions to the Y M C A to those given to churches is more useful a measure than simply the amount of money turned in. The ratio of hours spent in travel to persons met with is a fair index of the difficulty of two areas from the standpoint of concentration of population. The minutes per boy given by each member of the paid staff is a fair index of the dependence of the work upon the secretary rather than upon volunteer local leadership. Some would prefer it to be small, others large.

The yearbook of almost every religious organization attempts to record certain important facts which will be significant of other things. The number of members is recorded not so much because numbers are an end in themselves, but because numbers are one rough indication of the attainment of objective. The number of babies baptized tells something about the church and community, but it would tell more if stated in comparison with the birth rate for the area. The number of persons going into the ministry or mission field in proportion to college population tells something about the effectiveness of certain types of influence in a given college. Perhaps when the proportion is large, some may regard the work as less desirable. That does not, however, destroy the value or the importance of the measurement. The aim of such tests is to reveal the state of affairs, rather than to pass judgment upon it. If numbers or dollars, or hours have been unduly stressed in judging the efficiency of religious work, that is an indication of the abuse of the measures, rather than an indication of the meaninglessness of such figures. Interpretation must proceed cautiously with many items of countable information. Several causes and several results must usually be stated as possible meanings of the data. Some will hold true in some cases, some in others. A high number of conversions

reported may be due to unusually good work, or to a low standard, or to lack of records plus high self-esteem, or to one of a number of other causes. Which cause is most likely to have operated in a given case can often be determined from other data. The point of emphasis here, however, is that the tendency toward tests of religious educational endeavors does not wholeheartedly support the impatience of some with statistical reports. It is urged that the reports be continued, that endeavor be made to secure items more revealing and more worth counting, and that when secured the items be carefully interpreted and not left as supposed indications of a success they do not necessarily demonstrate.

a. Methods of Collecting Objective Data

The method of gathering objective data often presents serious problems. Reports to higher officials in an organization are apt to be colored by the standards by which the reporter thinks the work will be judged. Reports within religious organizations are notoriously erroneous. Religious workers have not included statistical carelessness and overstatement among the seven deadly sins.

If an investigator turns from reports to special questionnaires he runs afoul of a new set of difficulties. In the first place, people do not like to fill out questionnaires.[1] They do not feel that they should perform the investigations for somebody else. Often-

[1] Even scientists are no exception. The *Psychological Bulletin* for May, 1926, contains the following statement:

"At the twenty-third annual meeting of Experimental Psychologists held at the University of Pennsylvania, April 5 to 7, 1926, the following resolution was passed:

"Resolved, that this meeting deplores the increasing practice of collecting administrative or supposedly scientific data by way of questionnaires; and

"That the meeting deplores especially the practice under which graduate students undertake research by sending questionnaires to professional psychologists."

EDWIN G. BORING,
Harvard University
SAMUEL W. FERNBERGER,
University of Pennsylvania
HERBERT S. LANGFELD,
Princeton University
E. S. ROBINSON,
University of Chicago
E. B. TITCHENER,
Cornell University
R. S. WOODWORTH
Columbia University

times questionnaires ask for information which the persons addressed cannot readily give. Many persons of prominence might, if they chose, spend each day upon questionnaires, hours of time which represent inestimable value to humanity. For them to take time from their peculiar contribution to make out questionnaires would be immoral. Where this difficulty is not serious, where data are available and clerks can get it, there are still other objections. One is that questionnaires sent out by mail are apt to be returned by only a small per cent of the persons addressed. If several follow-ups are used, if the questionnaire is very simple, and the persons addressed are really interested in the problem or the person carrying on the investigation, the return will be much larger, of course. Still, it is very difficult to know just how those who did not reply would have answered. They are clearly different in their habits and interests from people who did answer. Perhaps they are less sympathetic with the approach suggested by the questionnaire. Perhaps they are reluctant to reveal conditions which might be thought of as unfavorable to them or their work. In any case, the answers represent a selected sample. They are not a chance sample. Formulae for reliability[1] must be used with care.

A third objection to questionnaires is that they are likely to be interpreted differently by different persons. A questionnaire recently observed by the writer asked persons to answer "Yes" or "No," to the question, "Are the relations between Y M C A and the churches distinctly friendly?" What did that mean? Did it mean most of the churches or all? Did it mean to some people, simply absence of vitriolic attacks? What it meant is suggested by the fact that 99.9 per cent of persons representing towns from 5,000 to 25,000 population in the United States answered "Yes." Even when it is recognized that questionnaires are opinion studies rather than collections of other facts, interpretation is still bedevilled by this subjectivity. What did the question suggest to the mind of the reader?[2]

Unless questionnaires employ controlled answers, requiring only checking or rating in some definite way, a fourth objection

[1] See pp. 237–243. [2] See note p. 58.

enters. Usually the person sending out the questionnaire has to interpret the answers. It is extraordinarily difficult to bring together in any sort of tabulation the uncontrolled answers to questions. The very categories in which the investigator chooses to tabulate his replies may bias the result. It may be that 20 per cent of those that answer mention a certain difficulty or result. Yet if everyone had been asked, directly, whether or not that difficulty or result were present in his case, the per cent might have been 90.

When several investigators, working independently, tabulate results from questionnaires, the results are often surprisingly different. One helpful method of attack on such a situation is to bring together a group of persons, some of whom represent each of the possible biases upon the question being studied, and to let that group, through discussion, decide the best meaning to be placed upon each answer.

The general rule in the light of these and other objections is never to use questionnaires when any other method of solving the problem is available and likely, at moderate cost, to yield useful data. They are labor-saving devices, as a rule, only to those who do not recognize their limitations. "If you want a thing well done, do it yourself," is excellently fitted to the research worker.

3. Paper and Pencil Tests

A type of objective measure, in addition to the collection of significant objective facts, is the paper and pencil test, now rapidly rising into popularity. Of course the old essay type of test and examination has long been regarded as of some value. College courses in Bible have examined as doggedly as did other departments. Occasional Sunday-school reviews have been written lessons, and teachers have usually viewed with dismay the results. It has often seemed more encouraging to stop testing and to trust that the Lord would not let all of the seed fall on barren ground. The new type examinations, however, have furnished a more intriguing plaything. They are less work to give and to score. They suggest interesting alternatives. Pupils usually enjoy them. Frequently it is possible to compare results

with other groups. They seem to yield a precision of result which makes it possible to startle the world with journalistic headlines.

The use of existing paper and pencil tests, and the composition of new tests are discussed more fully in the ensuing chapters. In general the comment may here be made that they are no more to be accepted without careful evaluation than are ratings or statistical reports.[1] New type tests sometimes have lower reliabilities than do essay examinations. Sometimes tests which are put out as tests of some character trait yield results almost totally independent of the trait after which the test has been named. Sometimes elaborate indirect tests are devised to get at matters which could better be answered by a simple question which all subjects would be willing and competent to answer.

a. Tests versus Questionnaires

Symonds has suggested a distinction between tests and questionnaires which it seems well to mention, although the line cannot be drawn with absolute precision. Tests measure what a subject can do, questionnaires indicate what he has chosen to do. Tests in this sense would present situations in which pupils could not make a showing better than their best. On a spelling test, an intelligence test, an arithmetic test, a Bible names test, or a memory repetition test, a subject may do his utmost and be outclassed by others. A questionnaire or question ballot on pacifism may be answered at the discretion of the subject so as to present one picture of himself, or a totally different picture of himself. A supercriminal might answer a questionnaire on the advisability of repealing the Ten Commandments in such a fashion that his paper could not be distinguished from the average of the poll of the Christian Endeavor societies.

Many tests endeavor to keep themselves out of the questionnaire class by some concealment of purpose. It might be that an attitude test given by itself would be very subject to human aspirations to make a good showing for a special purpose, whereas

[1] For further discussion of the problem of interpreting paper and pencil tests see WATSON, "Orient and Occident, an Opinion Study," *The Inquiry*, 129 E. 52 St., New York, 1927; and WATSON, "Character Tests," *Religious Education*, May, 1927.

if the same items were embodied in a long test, administered as an intelligence test, the idea of placing his best goods in the ethical show window would not occur to the subject. In other tests a very complicated scoring device makes "faking" difficult. In some tests a direct misstatement is made with reference to the purpose of the undertaking. There is a general recognition, however, that the tester should not place himself in any unethical relationship to the subject.

b. Ethical Problems in Testing

This is somewhat easier to accept as a general principle than to interpret specifically. Is it unethical to watch a subject when he does not know that he is being watched? Is it unethical for a physician, in the interests of diagnosis and treatment, to pretend that he is concerned about heart sounds when he is really trying to watch the muscular control of the subject? Is it unethical to use some equivocal or vague name for a test which will state the truth, but not the whole or exact truth? Is it ethically desirable to say nothing to pupils who are to be tested for cheating but to let them suppose they are taking an arithmetic test? At what point does the service of humanity enter as more than counterbalancing a minor inconvenience to some individual? If a few may give their lives in medicine that many may live, is it right to ask a few to suffer depreciation of character that many may benefit? These are but a few of the involved problems of the ethics of psychology and education, upon which relatively little exploration has been made in modern times. They open up a field of inquiry into which scientific controls in religious education will necessarily lead. At present no better guidance can be given than that which any group of persons of high intelligence and an unselfish devotion to religious ends may discover through discussion among themselves.[1]

4. Conduct Tests

Paper and pencil tests tend to measure limited reactions. They must not be interpreted as being more than somewhat indicative of other forms of conduct. Information is important.

[1] For one answer, see Watson, "Character Tests," *Religious Education*, May, 1927.

Ability to discriminate is important. Attitudes are important. The kinds of conduct (perseverance, carelessness, neatness, cheating, etc.) which can be measured by paper and pencil tests are important. These last measures verge over into a third type of measurement, namely, measurement of conduct reactions in natural or controlled situations.

Suppose a child is given an Easter lily to raise for the decoration of his Sunday-school room. Will he tend it carefully or let it die? Suppose he has chickens or pets for whom to provide? How many days does he forget to water them? Suppose he is taken for a hike in a public park where flowers are abundant but must not be picked. Will he or will he not pick them? Suppose he has an empty candy bag in his hand on a city street. Will he throw it in the gutter or carry it to the nearest waste box? Suppose he is asked to stay home from an interesting movie to help his mother with some work. What will his responses actually be? If he is given too much change at the store will he keep it or return it? These questions are among the best test questions for character that we can devise. The answer to them is not readily revealed by paper tests. The answer can be found by placing children in such situations and recording what happens.

Sometimes these situations are those occurring naturally day by day. These have the advantage of not seeming forced or unusual. There is nothing to put a subject on his guard. These advantages are balanced by two heavy disadvantages. One is that to test in this manner means that the person making the test must literally dog the footsteps of the pupil for several days or weeks. The other is that not all pupils encounter the same situations, or in so far as they do, the situations seem somewhat similar but are effectively different. Two pupils may both be seen to react to the situation "Little-brother-wants-my-toy." In one case, however, it may be that the toy is fragile, in the other case indestructible. It may be that one child has been taught not to give such things to little brother because they may hurt him, while the other toy has no such characteristics. Time of day, family atmosphere, the affection of the first child for his

toy, all these and other variables conspire to make such reactions very difficult of interpretation by an observer.

Hence a few attempts have been made to set up controlled situations. With these it is possible to place each one of a large number of children in very much the same situation. It is possible to see that each child responds not only to one situation in a week but to a dozen in a day, involving the same sort of choice. Yet the controlled conduct tests have their limitations also. It is usually impossible to control that important half of the total situation, which is within the body, thoughts, experiences, and feelings of the subject. Also, the controlled situations are apt to seem unnatural and forced. Coming into a room with a given examiner and being asked to do certain stunts must arouse active questioning in the mind of any intelligent child. Voelker, Cady, Raubenheimer, and others have used conduct tests in what seems to the writer too great faith that the pupils did not know what it was all about. When Terman found that the intellectually superior children in California were less likely to perform certain untrustworthy acts than are average children, it was not necessary to accept his interpretation (although the statement is probably true) that character and intelligence go together. It may be true that the bright children caught on and tried to make a good showing, the duller children more naïvely gave themselves away. When Voelker found that boys of eleven did not improve so much on the second test as did equally bright but more intellectually sophisticated boys of thirteen or fourteen, it was surely not necessary to assume as he did, that the second group had become better characters. There is some tendency at least for controlled conduct tests to become behavior questionnaires in the sense of Symond's distinction, since children can suit the answer to the interest of the moment.

5. Interview Methods

Interview methods may be considered as examples of any of the types of measurement so far considered. They may be considered as forming a basis for ratings, in which case, all the 10 precautions suggested in connection with ratings may well be borne in mind. Interviews may be considered as being the

same sort of thing which one finds in paper questionnaires and tests, carried on orally. There is very little difference in the principles which govern measurement when a child is asked on paper to mark his opinion about swiping apples, or the minister's sermons, and the principles which govern the same tests applied orally, in a conference. The oral situation does, of course, tend to be more difficult to control. The raised eyebrow of the examiner may deflect the course of an answer. The American Social Hygiene Bureau has found that written anonymous questionnaires on sex practices yield truer results than do interviews on the same matters with qualified psychiatrists. Dr. Hamilton, in the interests of accuracy in his interviews, wired the chairs to the floor so that positions of examiner and subject would not change too markedly.

Again, an interview may be regarded as a form of conduct observation. It may be of the natural uncontrolled type in which the person is encouraged to talk along about whatever interests him, or of the more careful, studied variety in which certain questions are asked in certain ways, with a view to observing just what the person will say or do under such circumstances. Stenographic notes or moving pictures may supplement ordinary observation in carefully controlled interviews. No general appraisal of interviews can be made. For certain purposes, carried on with proper controls, interviews afford the best measures. In many other cases interviews as measures, whatever may be their educative value, are less useful than other forms of rating, questionnaire, paper test, and conduct test.

6. Case Studies

One other approach to character measurement must be given some consideration. This is the method of making case records, case studies, or case histories. Case studies are difficult to evaluate because they range along a line extending at least from abbreviated collections of adjectives expressing the opinion of one individual based entirely on hearsay, to Boswell's "Life of Johnson." The distinctive advantage of a case history is that it recognizes the interrelationships of factors, and does not fall into the error of regarding an individual as made up of separate facets

like a diamond, each of which can be measured without much regard for the others. It also permits distinctions between important and unimportant, impressive and unimpressive, apparently significant and apparently insignificant traits. In the present state of knowledge it is the most comprehensive measure, if well done.

It should be recognized, however, that case histories are really combinations of the measures mentioned before. The bringing together of all available measures about individuals, and building these into a structure in which chronological sequence and the presuppositions of the investigator about causal relationships are the cement, constitutes a case history. It is hazardous to assume that case studies built upon originally unreliable measures will be very significant. Consider for a moment this quotation from a case history published by one of the best organizations in the country for the study of cases of personality adjustment.

> We have been interested in the study of two brothers. One at two years, is vivacious, active, and pugnacious, striking and slapping his brother or playmates to gain his ends and domineering over his brother two years older. He is responsive and friendly, readily drawing adults to him because of his cheeriness and air of bravado. The other, at four years, is quiet and retiring. He gives way at once to his little brother's demands, making no effort to assert his rights, is shy and unresponsive, and inclined to be fretful.

How does a reader know that one boy is vivacious and the other unresponsive? How was this determined? Is there any difference between such statements about activity, pugnacity, cheerfulness, shyness, fretfulness, and the sort of unchecked rating which evidence has shown to be wholly untrustworthy? Does building these loose opinions into a connected story make them "scientific"?

Case studies will be more useful in proportion as they are able to utilize refined measures for each element they take into consideration. Where fact records, such as date of birth, date of marriage, times arrested, years spent in attending school, and dollars of income are concerned, these are usually fairly reliable. Where they are not clearly established, the reservations should be

stated. A recent investigation in an orphanage showed that for a significant number of the children chronological age could not be determined. The birthday date on records was the date of commitment by the court. Where any aspect of knowledge, attitude, or behavior that can be tested is counted significant, the record should be entered as a test report. A teacher's statement that a boy is "bright enough but doesn't work" is no indication of the real state of affairs. Investigations have shown again and again that apart from accurate tests such statements are as likely to mislead as to furnish a good working basis. Yet the teacher probably has a better chance to know intellectual capacity than the ordinary religious or social worker has to observe the traits so readily jotted down as a "case study." Where it becomes necessary to include in a case record statements about general traits of character or personality which cannot be based on test scores or on reportable facts, then all the care should be exercised which is used in the best rating scales. The trait should be judged by experts in the particular kind of judgments involved, who have had more than casual opportunity to observe in a variety of circumstances. They should be judged by as many persons as possible, by persons not interested in proving any particular point, on the basis of a clearly defined and illustrated scale, preferably yielding a numerical rather than an adjective result, arranged so that each element in the total trait is analyzed out and given its proper weight, corrected for halo effect, bunching at one end of the scale, and for the degree of certainty of the judge. At best it should be remembered that ratings are usually based on one or two or three observations of the trait in question, be it cheerfulness, obedience, vivacity, fretfulness, or pugnacity. Even the shorter paper and pencil tests usually take into account fifty or a hundred elements, scattered over a wide range.

There is often a grave danger that the relationships which do appear in case studies were present first in the recorder's mind. Not infrequently it happens that one who believes mother fixation and broken engagements go together finds one causing the other, whereas someone else writing from the same facts would

find a feeling of physical inferiority causing both. The host of case studies now appearing in the important and little-developed field of emotional drives and wishes, needs careful scrutiny to separate observed facts from imputed relationships. By analogy, a case study surpasses the result from a battery of tests, in much the same fashion that a building surpasses a brick pile. Relationship is, indeed, of vital significance. But this does not justify building houses out of bean bags or personality studies on anyone's casual say so. The better case studies differentiate clearly between what is observed and the observer's interpretation of that behavior in terms of purposes, traits, motives, causes, and results. Such studies, especially when made by several persons over long periods of time, provide excellent clues for controlled experiments.

There is no real justification for antagonism between those whose interest is in tests and those who have found value in case studies. Each needs the modification and supplement offered by the other. Not controversy but active cooperation must be hoped for between those whose chief interest has been the reliable measurement of aspects of human behavior, and those whose major concern has been the meaningful interpretation of the total behavior in one individual.

EXERCISES

1. Indicate your opinion about each of the statements listed below by using a figure from the following scale:

Use 5 to mean surely completely true.
4 to mean probably true, or true in large degree.
3 to mean doubtful, don't know, about even.
2 to mean probably false, or false in large degree.
1 to mean surely, completely false.

____ *a.* A reliable test for courtesy could not be a measure of conformity.
____ *b.* Some qualities can never be measured in any way.
____ *c.* Tests are less fakable than questionnaires.
____ *d.* It is possible to carry on religious work with no measurement.
____ *e.* The experimenter should never deceive the subjects.
____ *f.* Experience is important in determining the usefulness of a rater.

____ *g.* Intimate friends are the best sources for ratings.
____ *h.* A rating scale which rates all persons on one trait is better than one which rates one person on all traits.
____ *i.* Reliability is more important than validity.
____ *j.* Persons who invent a new method are likely to be poor judges of its success.
____ *k.* The midmark is preferable to the average of a series of ratings of the same trait in the same person.
____ *l.* Specific traits are better rated than are general ones.
____ *m.* Certain features of physical appearance are good indices of character traits.
____ *n.* Score cards increase reliability.
____ *o.* The halo effect is indicated by low intercorrelations among traits.
____ *p.* The fact of certainty is more significant than its interpretation by the rater.
____ *q.* The reliability of a doctor's diagnosis would be higher than that of an intelligence test.
____ *r.* Rugg's evidence served to encourage many to use ratings instead of tests.
____ *s.* Questionnaire research is becoming more popular and desirable.
____ *t.* Controlled observation facilitates comparison over ordinary case records.

2. Suppose it were desired to measure the reverence produced in a given group of twelve-year-old boys by a given pageant. Indicate how each step in the improvement of crude observation would apply to such a situation.

3. What matters of objective fact would serve as some indication of the effectiveness of a denominational religious establishment upon the campus of a state university?

4. Find a case study in one of the following or other sources. Pick out five elements which seem to you to be objective and highly reliable, and five which you would question.

Judge Baker Foundation, Case Studies, Series 1
Sayles, "The Problem Child in School"
Anon., "Three Problem Children"
Pressey, "Mental Abnormality and Deficiency"
McDougall, "Outlines of Abnormal Psychology"
Files of *Mental Hygiene*

5. Prepare a brief questionnaire, designed to get at the facts upon some problem, which would meet all of the objections raised to questionnaires in this chapter. Note those you cannot avoid, if any.

CHAPTER IV

TESTS NOW AVAILABLE

1. The Growing Field

The measuring instruments available for those interested in religious education are rapidly increasing in number. Five years ago the tests which seemed to offer promise in the field of moral attitudes could have been counted on a speaker's fingers. Now the *Psychological Bulletin* plans for an annual summary, beginning with the survey by Mark May in the *Psychological Bulletin*, July, 1926. A cross-section of the stream of development at any moment is likely to be behind the times before it can be published. The following presentations are therefore made with apologies to those who shall have constructed more useful instruments before these words fall into their hands. Constant perusal of all the periodicals of education and psychology is the only method which may keep one generally abreast of the movement. Even then there will be developments here and there in religious organizations, and especially in graduate school research which will have escaped observation.

The difficulty is further increased because the interests of religious education are expanding. At one time, perhaps, Bible knowledge tests would have been sufficient. Today there is a widespread conviction that Biblical material is taught, not for its own sake, but in order to influence attitudes and behavior. Hence it is demanded that tests be applied to the real objective, the resulting ideals and conduct, rather than merely to the intermediate objective of Bible information acquired. Many would expand their concern beyond everyday morality into the wider concerns of social welfare. They would not care to pronounce upon an individual's religious education until his attitudes toward war, birth control, and extraterritoriality have been made clear. Still others would contend that no particular information or

conviction should be regarded as the goal for religious education. They would measure results in terms of the process of living which has been set up.

Tests to include their concerns must take principal account of fair-mindedness, willingness to think problems through seriously and without flinching, readiness to act upon the best evidence available, always remaining alert to see new points of view. If life is somehow constantly examining itself and enriching itself in the interests of the largest good which can be recognized, then some feel religious education to have been successful, whatever the information or lack of information. Recently an additional emphasis has been laid upon the proper adjustment of emotional drives, the integration of desires, resolution of conflicts, and balancing of personality. In all probability this will not mark the final addition to the concepts of religious educators regarding the real nature of their task. Yet each of these attitudes demands certain distinct and additional tests.

2. Intelligence and Vocational Tests

The field of intelligence tests will not be covered in this summary, not because it is of little significance to the experimenter in religious education but because it has been so frequently, ably, and completely done elsewhere. The most reliable measures of general intelligence for children are the Stanford Revision[1] of the the Binet-Simon scale and the Herring Revision[2] of the same scale. Each of these is a test which must be given to one child by one examiner, and occupies an interview of from 15 to 100 or more minutes. Special training in the technique of administration and interpretation is required. These examinations yield very high reliabilities, perhaps 0.98 or 0.99 and correlate very closely with one another. For group tests, the written examinations prepared by the Institute of Educational Research, those prepared by Thorndike, and those prepared by Thurstone yield

[1] Terman, "The Measurement of Intelligence," 362 pp., Houghton, Mifflin Company, Boston, 1916.

[2] Herring Revision of the Binet-Simon Tests, *Examination Manual:* Form A, 56 pp., The World Book Company, Yonkers, 1922.

very high reliabilities. It is hardly safe to rely on an individual intelligence quotient determined by such an examination of less than two hours' length. The long college entrance intelligence examinations are probably more reliable than are the ordinary Binet examinations. There are a large number of group intelligence tests adapted to easy administration in a single class period.[1] With primary children the Detroit or Pintner-Cunningham tests are often used. In the grade school the National Intelligence Tests, the Haggerty tests, the I. E. R. Tests of Mental Ability, and recently the McCall Multimental tests have found wide favor. In high school the Illinois, the Terman, and the Otis Self-administering tests are often used. In addition to these tests of abstract intelligence there are several tests of intelligence of a more mechanical sort, notably the Pintner-Paterson Performance Scale. Instead of problems to be worked out with pencil and paper, the ingenuity of the child is tested by blocks to be fitted into molds and puzzles to be put together. One of the best presentations of the present status of intelligence tests may be found in Thorndike's "The Measurement of Intelligence."[2]

Another group of tests which is of distinct importance to religious educators but which cannot be summarized here is the group of special vocational tests. Some of the more general tests of interest have been included, but no tests of ability such as the Stenquist tests of mechanical ability, which require a subject to fit pieces of machinery together and to recognize missing parts in pictures of simple machines. Toops[3] has published much

[1] The largest publishers of tests of mental capacity and school achievement are:
The World Book Company, Yonkers, N. Y.
The Public School Publishing Company, Bloomington, Ill.
The Bureau of Publications, Teachers College, Columbia University, New York, N. Y.
C. H. Stoelting and Company, 424 North Homan Ave., Chicago, Ill.

[2] Teachers College Bureau of Publications, 1927. See also PINTNER, RUDOLPH, "Intelligence Testing," Henry Holt and Co., New York, 1924.

[3] TOOPS, "Tests for Vocational Guidance of Children Thirteen to Sixteen," Teachers College, Columbia University, New York, 1923.
TOOPS, "Trade Tests in Education," Teachers College, Columbia University, New York, 1922.

useful material on such tests. Miss Manson's "Bibliography of Psychological Tests and Other Objective Measures in Industrial Personnel"[1] lists a number of studies made with tests. Kornhauser and Kingsbury[2] discuss the task of making and applying tests for business. A very useful bibliography on the entire field is issued by the Board for Vocational Education.[3]

3. Description of Tests Now Available

In the following pages a description is presented of tests which are now[4] available in published form. These can be bought and used by workers in religious education. In the case of each test, all available information has been gathered under the following heads:

Name of Test.
Publisher.
Price, in orders of 25 unless otherwise stated.
Purpose, as stated or as seems apparent.
Contents, with samples where space permits.
Administration, with reference to the age of group to which the test is suited, the time required for giving it, usually generously estimated, special conditions necessary to be observed in giving it.
Scoring Method used.
Data on *Standardization*, with reference to reliability, validity, norms, and certain other *Comments* or special references.

In each case the author or publisher of the test has been invited to supplement this information with any more recent evidence. It must be remembered that this is a cross-section of

[1] Manson, "Bibliography on Psychological Tests and Other Objective Measures in Industrial Personnel," Reprint and Circular Series of the Personnel Research Federation, 40 West Fortieth St., New York.

[2] Kornhauser, A. W., and Kingsbury, F. A. "Psychological Tests in Business," University of Chicago Press, Chicago, 1924.

[3] "Bibliography on Vocational Guidance," Federal Board for Vocational Education, *Bulletin* No. 66, Trade and Industrial Series No. 19, Revised Edition, Washington, D. C., December, 1925.

[4] Prepared June, 1926.

a moving stream, and that from time to time new tests will be developed and old tests better standardized.[1]

Name of Test: Biblical Knowledge Tests.
Publisher: M. T. Whitley, Teachers College, Columbia University, New York, N. Y.
Price: 6 cents each.
Purpose: Measurement of knowledge about Old and New Testament. Non-critical.
Contents: One form for Old Testament, two forms for New Testament, Multiple choice items.
Solomon was the son of _____ Samuel, Saul, David, Levi.
Administration: Age of group: ten years and over.
Time required: 30 minutes, each form.
Special conditions: Old Testament required timing of each section.
Scoring: Furnished with tests. Count number right.
Standardization:
Reliability.
Validity.
Norms (Old Testament Based on 120 children): Age 11–40
New Testament tests being standardized 14–46
now. 15–52
16–59
Other Comment: For later development, consult:
Professor M. T. Whitley,
Teachers College,
Columbia University,
New York, N. Y.

[1] The most comprehensive bibliographies at present are:
a. HARTSHORNE and MAY, "Objective Methods of Measuring Character," Pedagogical Seminary, March, 1925.
b. WATSON, "A Supplementary Review of Measures of Personality Traits," *Jour. of Educ. Psy.*, February, 1927.
c. MANSON, "A Bibliography of Tests of Character," published under the auspices of the National Research Council, 1926.
d. MAY and HARTSHORNE, "Character and Personality Tests," *Psy. Bull.*, July, 1926, and July, 1927.
e. WATSON, "Character Tests of 1926," *Vocational Guidance Magazine*, April, 1927.

Name of Test: Bogardus Social Distance Test.

Publisher:

Modification dealing with international feelings prepared by the Inquiry into the Christian Way of Life, 129 E. 52d St., New York City.

Price:

Purpose: To measure the extent to which persons of other races, nations, religions, or economic classes are welcomed into the fellowships of social life.

Contents: After the name of each race (in the racial test) or the corresponding elements in other tests, there are seven columns representing seven stages of intimacy. The first and closest is willingness to admit members of the race, as a class, not necessarily the best or worst members, to close kinship by marriage. Following that are the steps of admission to home, to employment, to street or neighborhood, to church, to citizenship in the United States, and admission into the country as an alien.

Administration: Age of group: Adult.

Time required: About 30 minutes.

Special conditions:

Scoring: A group profile may be developed, showing the social intimacy to which each of the races, nationalities, or classes will be admitted by the average person in the group. Individuals may be compared with the group.

Standardization:

Reliability.

Validity.

Norms.

See Journal of Applied Psychology, IX: 216 and Journal of Applied Sociology, May–June, 1925.

Other Comment:[1] For later developments consult:

Professor Emory H. Bogardus,
University of Southern California,
Los Angeles, Calif.

Name of Test: A Brief Test in Religious Education.

Publisher: Research Department, National Council of Y M C A, New York, N. Y.

Prices: 5 cents in groups of 25 or more.

Purpose: To give in brief time a fair idea of the religious and ethical development of groups in Sunday schools.

[1] See BOGARDUS, "Social Distance Test," *Jour. App. Psy.*, May-June, 1925.

Contents: 20 "Yes-No" questions on the Bible, God, Jesus, prayer, churches, etc.

Ten instances from everyday life in which the subject is asked to judge which of four or five alternatives would be the best thing to do or say.

Administration: Age of group: Over nine years of age.

Time required: About 15 minutes for eighth graders.

Special conditions:

Scoring: Based on boy's answers, obtained through the Y M C A.

Standardization:

Reliability: Self-correlation for individual — 0.45.

Validity: Agrees fairly well with leaders' estimates of boys in their own clubs.

Norms: Based on 3,000 replies. Perfect score — 100. Average boy, age 13, score — 0. Poorer than average below zero.

See *Program Papers* No. 5, Association Press, New York.

Other Comment: Mimeographed only.

Ethical test taken from C.E.I. Comprehensions.

For later developments consult:

Goodwin B. Watson
Teachers College,
Columbia University,
New York, N. Y.

Name of Test: The Brotmarkle Comparison Test.

Publisher: C. H. Stoelting and Company, 424 N. Homan Ave., Chicago, Ill.

Price: 6 cents in groups of 25. Manual 40 cents extra. Scoring guide 50 cents extra.

Purpose: To find the comparative content of moral concepts.

To show how nearly subjects agree with the average in the meanings they attach to certain words having ethical significance.

Contents: Seven seven-word tests. Extremes are located thus:

Purify _____ _____ _____ _____ _____ _____ _____ Corrupt

with the blanks between to be filled by the words cleanse, spoil, mend, injure, better, harm, reform.

Later subjects may alter words at extremes, write in other words, and list certain words believed expressive of his own character.

Administration: Age of group: College or high school.

Time required:

Special conditions:

Scoring: Deviation from norms based on 331 college students.

Standardization:

Reliability:

Validity:

Norms: Deviations in one class in first-year psychology:

V — 8–16	II — 28–34
IV — 18–20	I — 36–64
III — 22–26	

Other Comment: Discussion in *Journal of Applied Psychology*, 6: 235–242, 1922.

For later developments consult:

C. H. Stoelting and Company,
424 N. Homan Ave.,
Chicago, Ill.

Name of Test: Church School Examination Alpha. (A Revision of the Giles Sunday School Examination A)

Publisher: W. L. Hansen, Boston University School of Religious Education.

Price:

Purpose: To test knowledge of Biblical facts as presented by uniform lessons, and to test knowledge of ethical principles as in the Giles test, but free from the objections of the True-false form.

Contents: 25 Old Testament, 25 New Testament, 25 ethical elements.

1. Christ was born in:
   - ____ Jerusalem.
   - ____ Bethlehem.
   - ____ Nazareth.
   - ____ Capernaum.

53. Toleration is the opposite of:
   - ____ treason.
   - ____ coöperation.
   - ____ dishonesty.
   - ____ bigotry.

26. Adam and Eve were driven from Eden because:
   - ____ other people did not like them.
   - ____ they disobeyed God.
   - ____ the flood came.
   - ____ Cain killed Abel.

52. It is our duty concerning those who are the victims of injustice to
   - ____ mind our own business.
   - ____ tell them that they deserve it.
   - ____ have them arrested.
   - ____ help them.

Administration: Age of group: 10 and above.

Time required:

Special conditions:

Scoring:

Standardization:
Reliability:
Validity:
Norms:
Other Comment: For later development consult:
W. L. Hansen,
Boston University School of Religious Education,
Boston, Mass.

Name of Test: Colgate Emotional Hygiene Test.
Publisher: *The Hamilton Republican,* Hamilton, N. Y.
Price: Schedule B — psychoneurotic traits.
Schedule C — Introversion-extroversion.
$6 per hundred for either.
Purpose: To discover unhealthy emotional outlets and emotional types among college students. To help in vocational guidance and personality adjustment.
Contents: A series of elements like those in the Woodworth questionnaire, each followed by a graphic scale. Traits grouped according to classical clinical entities, such as psychasthenoid, schizoid, neurasthenoid, hysteroid, and introversion-extroversion.

Have you had trouble walking in the dark?	Never	Only when dark and stumbled	Occasionally	Every time
Have you preferred to be alone?	Always sought company	Enjoyed company but did not seek it		Always sought solitude
How have you been at selling things?	Delighted in convincing customers	Have sold many	Will take orders but will not sell	Avoided selling

Administration: Age of group: College, industry, business men.
Time required: About hour for both schedules combined.
Special conditions: No time limit. Self-administering.
Scoring: The present tests are the result of four years experimental work. They have been found useful in personnel work in plants and colleges. They may be taken individually or in groups. There is no time limit. As a rule about 30 minutes is required to complete each one.
The tests are scored by stencils which indicate each response in which the individual is "more so" than three-fourths of college students, that is, deviates from normal. A test requires about 80 seconds to score. Scoring stencils and percentile tables for men and women are supplied users of these forms.

All forms of the Colgate Emotional Hygiene Tests correlate zero with intelligence in the case of subjects of high school or college intelligence.

Personal Inventory C_2: for introversion. The reliability of this form is 0.85. The score on this correlates about 0.40 with scholarship. When the introvert score is combined with an intelligence score the coefficient of alienation in predicting scholarship is about doubled. The student more introvert than 75 per cent of his fellows has only one chance in 40 of being on probation, while the one less introvert than 75 per cent has one chance in 7.5.

The introvert is the student type. Lack of introversion has been found to be associated with success in selling.

In this form of The Colgate Emotional Hygiene Test the individual describes himself.

Personal Inventory C_3: for introversion. The individual under observation is described by those who know him *fairly* well, such as teachers, foremen, roommates, parents. The objectivity of this form is 0.85, the reliability 0.90.

The uses are the same as for Personal Inventory C_2 but this form is designed especially for use in industry, psychopathic patients, and in other cases where coöperation or sincerity may be open to question.

Personal Inventory B_2: for signs of emotional instability or unbalance. This edition of The Colgate Mental Hygiene Test has three sections. The first is devoted to neurasthenia, the second to schizophrenia, the last to psychasthenia. In this test the individual describes himself. The reliability is 0.88. This has no predictive relation with scholarship but is extremely useful in detecting those who deviate from normal in emotional outlets.

The tests are sold in bundles of 100 of A form for $6. Stencils and percentile tables are supplied all users.

Other Comment: See *Journal of Educational Psychology*, September, 1925
Journal of Applied Psychology, p. 293, September, 1925
Journal of Abnormal Psychology, July, 1925
Science, September 25, 1925
Journal of Abnormal Psychology, March, 1926

For later developments consult:

Prof. Donald Laird,
Colgate University,
Hamilton, N. Y.

Name of Test: Downey Will-temperament Test.

Publisher: The World Book Co., Yonkers-on-Hudson, New York, or 2126 Prairie Ave., Chicago, Ill.

Price: Group Test 6 cents in groups of 25. Manual 15 cents.

Purpose: To determine the temperamental traits of individuals through a series of motor reactions.

Contents: Tests based largely, although not entirely, on handwriting. Tests ask pupils to indicate adjectives which fit themselves, to write as rapidly as possible, as slowly as possible, to practice copying a model, to write with eyes shut and while counting taps. Pupil's reaction to direct contradiction is included.

Administration: Age of group: Fifteen and over.

Time required: 50 minutes, if examiner is competent.

Special conditions: Examiner should have preliminary apparatus such as stop watch, pencil, white card, envelope, etc.

Scoring: A series of samples are provided in the manual of directions. Scoring is time consuming. Yields a profile showing "speed of movement, freedom from load, flexibility, speed of decision, motor impulsion, reaction to contradiction, reaction to opposition, finality of judgment, motor inhibition, interest in detail, coördination, and volitional perseveration."

Standardization:

Reliability: Mean for many investigations between 0.05 and 0.40 Downey's results indicate higher reliability, especially for tests of speed of movement, motor inhibition, and motor impulsion.

Validity: Correlation with ratings from −0.65 to 0.54, mean between 0.00 and 0.25.

Norms: Significant with some individuals.

Establish percentiles within group.

Other Comment: Individual test probably more valuable than group test. Best presentations found in Downey, "The Will Temperament and its Testing" (The World Book Co.); May, "The Present Status of the Will-temperament Tests," *Journal of Applied Psychology*, January, 1925; Uhrbrock, "The Downey Will-temperament Test."

For later developments consult:

Prof. June Downey,
Department of Psychology,
University of Wyoming,
Laramie, Wyo.

Further discussion may be found in:

Bryant, E. K. — "The Will-profile of Delinquent Boys," *Journal of Delinquency*, pp. 294–309, **6** (1921).

Bryant, E. K. — "Delinquents and Non-delinquents on the Will-temperament Test," *Journal of Delinquency*, pp. 46–63, **8** (1923).

Horskovitz, M. J. — "A Test of the Downey Will-temperament Tests," *Journal of Applied Psychology*, VIII, pp. 76–88.

MAY, M. A. — "The Present Status of the Will-temperament Tests"; *Journal of Applied Psychology*, IX, pp. 29–52.

REAM, N. J. — "Group Will-temperament Tests," *Journal of Educational Psychology*, pp. 6–12, **13** (1922).

RUCH, G. E. and DEL MANZO, N. C. — "The Downey Will-temperament Group Test; Analysis of Reliability and Validity," *Journal of Applied Psychology*, pp. 65–76, **7** (1923).

WIRES, EMILY. — "The Downey Will-temperament Profile," *Journal of Abnormal Psychology and Social Psychology*, pp. 416–441, XX (1926).

Name of Test: Emotional History Record.

Publisher: Goodwin B. Watson, and J. O. Chassell, Teachers College, Columbia University, New York, N. Y.

Price: 10 cents each for any number. Used for experimental purposes only. Now being replaced by Chassell's "Experience Variables."

Purpose: To reveal students needing help in emotional adjustment; to study the relationship between past experiences and present symptoms.

Contents:

Part I: A modification of the Woodworth-Mathew Questionnaire, asking about present neurotic symptoms.

Part II: A rating of the interest possessed by self, average person, and ideal person, in each of 50 lines of activity.

Parts III and IV: Records of past experience which may have influenced emotional make-up, *e.g.*, health of parents, favored children in family, success in school, attitude toward other children, home atmosphere, etc. Largely controlled responses.

Administration: Age of group: Adult.

Time required: Two hours.

Special conditions: Should not be used except with advice on guidance from a capable psychologist or psychiatrist.

Scoring: Yields score for number and significance of answers indicating problems. Also scores for self-esteem, self-ideal conflict, difficulty of adjustment in present group. Scoring system in process of change. Best used as picture of present status of individual.

Standardization:

Reliability:

Validity: Checked against 100 interviews and found wanting at many points. In process of revision. Part II best.

Norms: Based on students in various schools of Columbia University.

Other Comment: Likely to be most useful to the scientific student of indi-

vidual problems and emotional difficulties. Now being revised and reprinted under the title "Experience Variables."

For later developments consult:

Goodwin B. Watson and J. O. Chassel,
Teachers College,
Columbia University,
New York, N. Y.

Name of Test: Fernald Achievement Capacity Test No. 19,433.

Publisher: C. H. Stoelting and Company, 424 N. Homan Ave., Chicago, Ill.

Price: $146.

Purpose: To test that function of the mind called will, persistency, determination, pluck, or spunk in terms of muscle fatigue in units of time.

Contents: A simple device for visualizing the degree of elevation of the subject's heels while he supports himself with the heels off the floor. Equipped with electric bell to warn when heels sink to floor level. The device consists of a plate upon which subject stands, facing an upright erected to a height of about 5½ feet. Pivoted on the plate is a stiff wooden lever with a light cross-bar at the end of the short arm on which the subject's heels rest and connected at the other end by a thin wire to the short arm of a 25-centimeter needle, pivoted at a point near the top of the upright to oscillate in a vertical plane behind a suitable dial. This apparatus is positive and delicate and visualizes even the involuntary muscular tremors communicated by the heels of the fatigued subject.

Fatigue is rapidly induced without harmful results by requiring the subject to stand with heels ¼ inch off the floor. The muscles fatigued are those whose strength and development correspond to the body weight, that is, the muscles used to support and carry the weight of the body. Previous training plays only a nominal part in this test.

Scoring: Timed noted to seconds. The average norm about 50 minutes seems to establish the fact that in applying the test, any subject who stands longer may be advised to rest as he has demonstrated normal capacity in this test. In evaluating this test not only will the wide variation between the lowest and highest scores in both reformatory and norms groups be noted (3¾ to 52¾ minutes and 12 minutes to 2½ hours, respectively); but also the marked disparity between the median and average of the two groups; *i.e.*, about 35 minutes, a difference twice as great as the reformatory group average.

Other Comments: See "Defective Delinquents Class, Differentiating Tests," *American Journal of Insanity*, Vol. XIII, No. IV, April, 1912.

Name of Test: Fernald Ethical Discrimination Test No. 36,035.

Publisher: C. H. Stoelting and Company, 424 N. Homan Ave., Chicago, Ill.

Price: 40 cents a set, single copy.

Purpose: Some criterion of the moral work of a subject is essential to a proper understanding of his mentality and to a correct judgment of the degree of his responsibility; and some insight may be gained by this test into the intellectual and moral "manner of the man" from a study of his arrangement in the order of their worth of ethical entities of varying degrees of gravity or worth.

Contents: One large card and ten slips of cardboard, containing offenses to be arranged in a series from least to greatest in order of their gravity.

Administration: Age of group:

Time required:

Special conditions: Here are ten offenses to be arranged in the order of their gravity. Read them all over and find the one of least gravity or consequence and place it opposite 1. When you find the worst one place it opposite 10 in much the same way you were asked to arrange the weighted boxes. Place at 2 the offense which is next worse than No. 1, then the one of next greater gravity, etc. Take all the time you need. This test is not timed. When you are satisfied with your work, let me know. Use is made of the card device described on page 536 of the *American Journal of Insanity*, Vol. LXII, No. 4, April, 1912. See scoring slips No. 19,014 at 25 cents.

Standardization:

Reliability:

Validity:

Norms:

Other Comment: For later developments consult:
C. H. Stoelting and Company,
424 N. Homan Ave.,
Chicago, Ill.

Name of Test: Fernald Ethical Perception Test No. 27,105.

Publisher: C. H. Stoelting and Company, 424 N. Homan Ave., Chicago, Ill.

Price: 15 cents each.

Purpose: The determination of the question of the possession of some worthy degree of a knowledge of right and wrong is an element of some importance in the problem of the degree of individual responsibility. Tests of this knowledge, however, are purely intelligence

tests and a knowledge of right and wrong *per se* is small safeguard against wrongdoing.

Contents: A card containing ten ethical questions, seven of which are to be answered by "Yes" or "No," and three by taking one or the other horn of a dilemma.

Administration: Age of group:

Time required:

Special conditions: The subject is given a sheet of paper bearing the letters A to J on the left margin and is asked to write his answers opposite the letters corresponding to the lettering of the questions.

Scoring: For a comparative study of the value of each question a record of the answers to each is kept. For an evaluation of the work of each subject in this test the percentage of correct answers is noted.

Standardization:

Reliability:

Validity:

Norms: See *American Journal of Insanity*, Vol. LXIII, No. 4, April, 1912.

Other Comment: For later developments consult:

C. H. Stoelting and Company,
424 N. Homan Ave.,
Chicago, Ill.

Name of Test: Freyd's Occupational Interest Blanks.

Publisher: C. H. Stoelting and Company, 424 N. Homan Ave., Chicago, Ill.

Price: 4 cents each in orders of 25.

Purpose:

Contents: List of 80 occupations for men (or 67 for women), with a chance to register any of five attitudes toward each.

Administration: Age of group:

Time required:

Special conditions: Separate blanks for men and women.

Scoring:

Standardization:

Reliability:

Validity:

Norms:

Other Comment: Discussion in *Journal of Applied Psychology*, pp. 243–254, September, 1922.

Journal of Personnel Research, pp. 319–328, October, 1922.

See also Freyd, "The Personalities of the Socially and Mechanically Inclined."

Psychological Monographs, No. 151, 1924.
Psychological Review Publishing Company, Princeton, N. J.
Being enlarged and modified by Cowdery and Strong.
See *Journal of Personnel Research*, August, 1926, September, 1926, October, 1926.
For later developments consult:

Dr. Max Freyd,
Personnel Research Federation,
40 West 40th St.,
New York, N. Y.

Name of Test: Giles Sunday-school Examination A.
Publisher: Boston University School of Religious Education, Boston, Mass.
Price: Now out of print.
Purpose: To measure knowledge of the type of material included in the uniform lessons. Also to get an expression of judgment on ethical problems within children's ordinary experience.
Contents: 25 Old Testament, 25 New Testament, 25 ethical elements.

1. Christ was born in Jerusalem *True* *False*
26. Adam and Eve were driven from Eden because they disobeyed God *True* *False*
52. It is our duty to help those who are the victims of injustice *True* *False*
53. Toleration is the opposite of cooperation *True* *False*

Administration: Age of group: 10 and above.
Time required: 20 minutes.
Special conditions: Announce at 18 minutes that 2 more minutes remain.
Scoring:
Standardization:
Reliability: Correlation between New Testament and Old Testament is 0.50.
Validity:
Norms: Given in Indiana Survey, Vol. II, p. 392, for over 2,000 pupils in 24 schools. Pp. 388 *ff.*
Order of difficulty of statements given, pp. 399 and 400.
Other Comment: See Indiana Survey of Religious Education (Doran), Vol. II, p. 382 *ff.*
For later developments consult:

Prof. W. L. Hansen,
Boston University School of Religious Education,
Boston, Mass.

Name of Test: Hart Personnel Assayer.

Publisher: Iowa Child Welfare Research Station, University of Iowa, Iowa City, Iowa.

Price: 5 cents each.

Purpose: To give an array of information about a subject's interests and attitudes which might be helpful in vocational guidance. More generally, to differentiate certain types of people from other types.

Contents: Nineteen lists of 15 items each, to be marked "Yes" if liked, and "No" if disliked. Then the five things about which one feels most strongly are underlined, and the one thing which arouses most feeling is double underlined.

Items contain: mice, suck, faith, sneak, sermon, triumph, ministry, Rembrandt, successful, read editorials, protest, Einstein, hospitality, read about murders, be scolded, succeed in business, work for social justice, prohibit smoking tobacco, more interest in local politics, etc.

True-false statements, about which degree of feeling is indicated in the same way. Examples:

"It is wicked to teach that there are things in the Bible that are not true."

"Our country right or wrong" is a noble sentiment.

Administration: Age of group: Over 12.

Time required: 40 minutes or more.

Special conditions: No time limit. Can be done at leisure.

Scoring: Author has list of words which seem to indicate interest in money making, culture, religion, internationalism, etc.

Standardization:

Reliability: Still in experimental stage.

Validity: Shuttleworth reports 0.90 correlation with ratings of fraternity brothers on money-mindedness.

Norms: Has been shown to differentiate children rated as honest from those rated as dishonest, also to distinguish delinquents.

Other Comment: See "Can We Measure Results in Teaching Social Science?" *Proceedings* of the Nineteenth Annual Meeting of the American Sociological Society, Chicago, 1924.

For later developments consult:

Prof. Hornell Hart,
Bryn Mawr, Pa.

Name of Test: Hart Test of Social Attitudes and Interests.

Publisher: Iowa Child Welfare Research Station, University of Iowa, Iowa City, Iowa.

Price: 5 cents each.

Purpose: To show the predominant likes and dislikes, attitudes, and points of view.

Contents: Mark like (+) or dislike (−) for:

Be lonely.
Hear symphonies.
Discover truth.
Vulgar shows.
Death of a near relative.
Protest against social injustice.
Commune with the Great Spirit.
Better pay for school teachers.
Repress gambling.

Administration: Age of group: 12 or above.

Time required: 40 minutes.

Special conditions: Can be done at leisure. No time limit.

Scoring: Each year can develop his own system, depending on his interest. Author has typewritten lists of the words he counts indicative of interest in formal religion, interest in the spirit of religion, interest in culture, in money making, etc.

Standardization:

Reliability: Depends on list.

Validity: Uncertain.

Norms: None.

Other Comment: For later developments see:

Prof. Hornell Hart,
Bryn Mawr, Pa.

Name of Test: Interest Analysis.

Publisher: Personnel Research Federation, 40 West 40th St., New York City.

Price: At present being modified by Professor Strong, Stanford University.

Purpose: To discover occupational interests.

Contents: Part I: 70 occupations, to respond with five degrees of liking.

Part II: Like or dislike for such elements as: Fat men, blondes, people with gold teeth, progressive people, optimists, spendthrifts, unmoral people, Southerners, golf, checkers, summer resorts, detective stories, Charlie Chaplin, living in the country, penmanship, civics, mathematics, etc.

Part III: Belief in superstition. About 20 questions on emotional habits.

Part IV: Pressey X–O list, to cross out those which are unpleasant.

Administration: Age of group: High school or over.
Time required: About 50 minutes.
Special conditions:
Scoring:
Standardization:
Reliability:
Validity:
Norms:
Other Comment: See *Journal of Applied Psychology*, XVI, p. 243, 1922.

Name of Test: Kent-Rosanoff Free Association Test.
Publisher: C. H. Stoelting and Company, 424 Homan Ave., Chicago, Ill.
Also, in Rosanoff's "Manual of Psychiatry," published by John Wiley and Sons, Inc.
Price: Stoelting, $1.50 per hundred. Description, $1.15. Record blanks, 80 cents per hundred.
Purpose: To reveal abnormalities of emotional reactions.
Contents: Words to which the frequency of each of the common responses is known.
Administration: Age of group: Norms based on adults.
Time required:
Special conditions: Given to individual. Note with stop watch the time for each reaction.
Scoring: Unusual reaction time indicates conflict.
Idiosyncrasy is indicated by responses of an unusual character, as indicated in the norms.
Standardization:
Reliability:
Validity:
Norms: Based on 1,000 adult responses, giving frequency of each association.
Other Comment:

Name of Test: Kohs' Ethical Discrimination Test.
Publisher: C. H. Stoelting and Company, 424 N. Homan Ave., Chicago, Ill.
Price: 27 cents each in orders of 25. Manual 65 cents extra.
Purpose: To measure significant ethical knowledge and ability to make essential moral judgments, interpretations, and decisions.

Contents: Exercise 1: Social Relations: What would you do when playmate hits you without meaning to? If find man just hung himself? If man makes a million dollars?

Exercise 2: Moral Judgment: Which is the worst: fighting, killing, hating, quarreling, hurting?

Exercise 3: Proverbs: Don't count your chickens before they're hatched means _____.

Exercise 4: Definitions of moral terms: good means dirty, right, break, or bad; tergiversation means operating, arguing, hardening, backsliding.

Exercise 5: Offense Evaluation: For such offenses as bathing, bigamy, deafness, lynching, etc., should one be praised, ignored, scolded, put in jail, in prison, or killed?

Exercise 6: Moral Problems: You should not throw hot water on a cat because: you only waste the water, hot water hurts the cat, cats bathe in cold water.

Administration: Age of group: 11 to adult.

Time required: 20 minutes.

Special conditions: Each exercise must be carefully timed.

Scoring: Directions in *Manual.* Probably author's judgment as to best answer.

Standardization:

Reliability:

Validity:

Norms: Based on barely 100 cases:

92–100 Average Adult

86–91 Sub-average Adult

Tentative classification:

68–85 Inadequate

43–67 Moral deficients

Other Comment: Limited study shows that it correlates with intelligence higher than with ratings on ethical traits.

For later developments consult:

Dr. Samuel C. Kohs,
Executive Director,
Jewish Federation,
732 14th St.,
Oakland, Calif.

Name of Test: Laycock Test of Biblical Information.

Publisher: University of Alberta Bookstore, Edmonton, Alberta, Canada.

Price: 5 cents each in orders of 25.

Purpose: To test knowledge of statements made in the Bible, without regard for creed, church, mental, or moral ability.

Contents: Test 1: The ark was built by: Noah, Isaac, Adam, Abraham.

Test 2: The Lord's Prayer contains: Father forgive them for they

know not what they do; Father, I thank thee that thou hast heard me; Father, keep them in thy name that they may be one; Thy will be done on earth as it is in heaven.

Test 3: Paul was beheaded at Jerusalem: True, False.

Test 4: The Thirteenth Chapter of I Corinthians teaches: Love, Duty, Honor, Justice.

Test 5: The Book of Revelation was written from: Jerusalem, Patmos, Rome, Athens.

Test 6: The Book of Proverbs is: a book of war, a book of history, a book of prophecy, a book of practical wisdom.

Test 7: One of the twelve disciples was: Mark, Matthew, Paul, Luke.

Administration: Age of group: 12 to 16.

Time required:

Special conditions: Each section must be timed. Intelligent adults finish readily before the limit.

Scoring: Score sheet furnished.

Standardization:

Reliability:

Validity:

Norms: Based on 1,115 cases in Canada, Mean 29.3, S. D. 16.6

Age:	12	13	14	15	16	Girls 30, Boys 28
Approximate mean:	30	30	30	27	25	

Other Comment: See *Journal of Educational Psychology*, May, 1925, p. 329.

For later developments consult:

Professor S. R. Laycock,
University of Alberta,
Edmonton,
Alberta, Can.

Name of Test: Lundholm Emotional Cross-out Tests.

Publisher: McLean Hospital, Waverly, Mass.

Price:

Purpose: "The purpose of the test is to find out if there is a standard reaction to certain mental content in the human individual, and if there are sex differences in such standard response. Furthermore, the test aims at studying the influence of social inhibitions on the response to the test. We have assumed that the social inhibitions would work in any one individual instructed to put his full name to the test. Consequently we have given out the test in two forms, one in which the subject is given this instruction, and another in which he is told not to put his name on the test. If the standardization is possible we intend to carry out the test on insane patients,

particularly cases of dementia praecox, in order to see how they might deviate in their response from the normal. We have reason to anticipate the possibility that the response to the test might be used as a complex indicator in the mentally deranged.

"We have until now made tests on about 1,000 people, 800 college girls and 200 college boys. Age of the girls 17 to 25 years. The test has been given both as group and as individual tests.

"The validity of the tests cannot be determined as the complete analysis has not yet been carried out."

For later developments consult:

Dr. Helge Lundholm,
McLean Hospital,
Waverly, Mass.

Name of Test: Miner's Analysis of Work Interest Blank.

Publisher: C. H. Stoelting and Company, 424 N. Homan Ave., Chicago, Ill.

Price: 9 cents each in orders of 25.

Purpose: To help discover special interests and abilities by suggesting how to observe one's own likes and dislikes.

Contents: Twenty-eight paired contrasts, such as:

indoor — outdoor
skilled hand work — skilled heavy work
slow movements — rapid movements
broad planning — attention to details
less responsibility — more responsibility
organizing people — constructing things
working by yourself — working with others
doing the same thing — wide variety of work
regular time for work — irregular time for work
work in one town — work requiring traveling
welfare work — taking part in entertainments

Administration: Age of group: High school or above.

Time required:

Special conditions:

Scoring:

Standardization:

Reliability: Above 0.70 for all interests but one.

Validity: Separates women taking secretarial course from those taking home economics, men interested in machine construction. (See attached sheet.)

Norms: Based largely on 800 students at University of Kentucky. Being extended.

Other Comment: See *Journal of Educational Psychology*, May, 1926.
Journal of Educational Research, April, 1922.
For later developments consult:
Prof. J. B. Miner,
University of Kentucky,
Lexington, Ky.

Name of Test: Multiple-choice Test of Religious Ideas.
Publisher: Boston University School of Religious Education, Boston, Mass.
Price:
Purpose: To ascertain what religious ideas individuals, whether children or adults, in church school or elsewhere, actually have.
Contents: Check the five best out of fifteen suggested answers to the questions:

1. What is the purpose of the church?
2. Why should we study the Bible?
3. Why should we pray?
4. How do you think of Jesus?
5. How do you think of God?
6. How do you think of the Holy Spirit?
7. What does it mean to be a Christian?
8. How does one become a Christian?
9. What is sin?
10. What do you think happens after death?

Administration: Age of group: "Fair facility in reading."
Time required:
Special conditions:
Scoring: Based on answers of specialists. Table given in Indiana Survey, Vol. II, p. 446.
Standardization:
Reliability:
Validity: Criticized by religious leaders and workers with tests.
Norms:
Other Comment: See "Indiana Survey of Religious Education" (Doran), Vol. II, pp. 430–450.

Name of Test: Pressey X–O Tests for Investigating the Emotions.
Publisher: C. H. Stoelting and Company, 424 N. Homan Ave., Chicago, Ill.
Price: 4 cents each in orders of 25. Manual 15 cents extra.
Purpose: To reveal individual differences in emotional and affective make-up, associational tendencies, moral sentiments, and anxiety tendencies.

Contents: Part I: 125 words such as disgust, fear, sex, suspicion, aunt, etc., for the subject to cross out those he dislikes.

Part II: List of associations: Choose the word among the following most closely connected with dream: floating, heart, beautiful, bicycle, manhood, etc.

Part III: Cross out the things for which people should be blamed: begging, swearing, smoking, flirting, etc.

Part IV: Cross out things about which have worried: injustice, noise, self-consciousness, germs, discouragement, etc. Encircle the words most strongly affective. Words include response of fear, sex, disgust, hypochondriacal nature, as well as jokes.

600 elements.

Administration: Age of group:

Time required: No time limit, about 50 minutes.

Special conditions: None.

Scoring: Affectivity: Relation of total numbers of words crossed out to norm for college students, etc.

Idiosyncrasy: Relation of number of words encircled which are not usually encircled, to the norm of such deviations.

Standardization:

Reliability:

Validity: Shown to be useful in distinguishing certain groups. Correlation of 0.43 with college grades. (See attached sheet.)

Norms: Based on 114 college students.

	Affectivity	Idiosyncrasy
Q_1	200	42
N	230	47
Q_3	260	52

Other Comment: Form A contains the original test. Form B is expurgated and supposedly more suitable for children.

See *Journal of Abnormal Psychology*, April, 1921.

Chambers uses parts of Form B as a measure of emotional maturity.

See *Pedagogy Seminary*, Vol. 32, No. 4, 1925.

Journal of Abnormal and Social Psychology, p. 303, October, 1925.

Validity: Olsen found idiosyncrasy significant in Manhattan State Hospital Cases, the average for 24 being 54 per cent above the norm, while average affectivity was 5 per cent below norm. Landis found idiosyncrasy better for predicting general emotionality.

Other Comments: For later developments consult:

Prof. S. L. Pressey,
Ohio State University,
Columbus, Ohio.

Name of Test: Test of Racial Attitudes.
Publisher: Goodwin B. Watson, Teachers College, Columbia University, New York, N. Y.
Price: 5 cents.
Purpose: To measure changes in attitudes toward persons of other races.
Contents: (36 elements)

All, Most, Many, Few, No ____ Jews will try to get the best of a bargain even if they have to cheat to do so.
All, Most, Many, Few, No ____ colleges should admit pupils without regard for race distinctions.
All, Most, Many, Few, No ____ persons of one race should be permitted, with mutual consent, to marry a person of some other race.

Administration: Age of group: 12 and over.
Time required: 20 minutes.
Special conditions: No time limit.
Scoring: (*a*) Extremism: Total number of "all" and "no" answers.
(*b*) Viewpoint: Items are classified as "liberal" or "conservative," the former tending toward racial equality, the latter toward preservation of race distinctions and barriers.
Standardization:
Reliability: 0.82 for extremism using split halves.
Validity: Showed definite change in a six-week class in a church.
Norms:
Other Comment: Mimeographed only.
See description in Watson, "The Measurement of Fair-mindedness," pp. 40–42.
For later developments consult:

Goodwin B. Watson,
Teachers College,
Columbia University,
New York, N. Y.

Name of Test: Test of Social Relations. (Ream)
Publisher: Carnegie Institute of Technology, Pittsburgh, Pa.
Price:
Purpose: To reveal interests in terms of acquaintance with the vocabulary of certain activities, such as baseball, poker, hymns, popular music, literature, etc.
Contents:

Administration: Age of group:
Time required:
Special conditions:
Scoring:
Standardization:
Reliability: Correlation of each half with total is 0.97.
Validity: Correlation with intelligence is 0.60.
Norms:
Other Comment: See *Journal of Applied Psychology*, p. 69, Vol. VI, 1922.
For later developments consult:

Dr. M. Jay Ream,
750 Broad St.,
Newark, N. J.

This test proved of little value in determining natural aptitude for sales work.

Name of Test: A Survey of Public Opinion on some Religious and Economic Issues.

Publisher: Bureau of Publications, Teachers College, Columbia University, New York, N. Y.

Price: Test with directions for finding gross score, 6 cents in orders of 25 or more. Test with analytical score sheets, 12 cents in orders of 25 or more. Mimeographed Tables, 10 cents.

Purpose: To measure fair-mindedness as contrasted with prejudice. To show the amount of prejudice in agreement with certain typical points of view, conservative and radical, within religious, moral, and economic questions.

Contents: Form A: Cross out disagreeable words in list containing Bolshevik, Sunday Blue Laws, Capitalist, Virgin Mary, etc.

Form B: Mark on + 2 to − 2 scale statements such as "Churches are more sympathetic with capital than with labor."

Form C: Draw possible inferences from such facts as "2 per cent of the population control 60 per cent of the wealth."

Form D: Pass judgment on instances such as the recognition of the Soviet, the healings at Lourdes.

Form E: Rate as strong or weak, arguments on child labor, divorce, the Ku Klux Klan.

Form F: Indicate All, Most, Many, Few, or No, as the best beginning for such statements as:

"_____ church members have private booze stocks in their cellars.
_____ strikes have been due to the laziness or greed of the workingmen or their leaders."

Administration: Age of group: Seniors in high school or above.
Time required: 40–60 minutes.
Special conditions: None. Can be done outside class, if desirable. Fact that it is a test of prejudice must not be mentioned.

Scoring: Directions given with test.
Gross Score: General level of prejudice obtained by finding per cent of possible deviations from fair-mindedness. Deviations consist in crossing out words, making extreme statements, inability to see the other side of an argument. Analytical score shows directions of prejudices.

Standardization:
Reliability: Gross score 0.96
Analytical scores: 0.60 to 0.88
Validity: Based on face value, intercorrelations, select individuals and groups (80 per cent agreement with these).
Norms: Normal school students..... Gross score 25 per cent.
Methodist ministers..... Gross score 30 per cent.
Liberal theological students..... Gross score 18 per cent.
Very prejudiced persons..... Gross score 50 per cent.
Very fair-minded persons..... Gross score 6 per cent.
Table of additional norms given in manual of directions.

Other Comment: Complete discussion, together with suggestions for measuring race prejudice and other such biases is found in the author's "The Measurement of Fair Mindedness," published by the Bureau of Publications, Teachers College, Columbia University. Brief discussion, *Industrial Psychology*, February, 1927.
For later developments consult:
Goodwin B. Watson, Teachers College,
Columbia University,
New York, N. Y.

Name of Test: Union Test of Ethical Judgment.

Publisher: Department of Religious Education, Union Theological Seminary, 3041 Broadway, New York, N. Y.

Price: 10 cents each, 5 cents in sets of 25 or more. Edition now exhausted.

Purpose: To measure the development of enlightened social and ethical standards, especially for religious educational groups.

Contents: Questions largely of the Yes-No or Excellent, Fair, Poor, type, dealing with choosing courses in school, finding life work. Prohibition, use of money, sweatshop labor, economic justice, the Kingdom of God on earth, gambling, charity, success in life, activities for a Sunday school class, leadership, fouling in a game, questionable

recreation, control of temper, petting, quarrels, true greatness, attitude toward parents, attitude toward other children, gifts, health, mental development, social growth, and other everyday acts.

Administration: Age of group: Form I: Grades III and over. Form II: high school and adults.

Time required: 50 minutes.

Special conditions: No time limit.

Scoring: Scores based on answers of 300 pupils (mainly in Union School of Religion) and theological students at Union and the University of Chicago. Score sheets furnished.

Standardization:

Reliability: 0.92 (Form II).

Validity: Correlation of only 0.23 with intelligence.

Norms: Being gathered.

Other Comment: Often leads to excellent discussion.

For later developments consult:

Goodwin B. Watson,
Teachers College,
Columbia University,
New York, N. Y.

Name of Test: Union Test of Religious Ideas.

Publisher: Department of Religious Education, Union Theological Seminary, 3041 Broadway, New York, N. Y.

Price: 10 cents each, 5 cents in sets of 25 or more. Edition now exhausted.

Purpose: To measure intellectual factors in religious development.

Contents: Questions, mainly answered "yes" or "no," on ideas about God, Jesus, Prayer, The Church, Bible Interpretation, Other Modern Religions. Completion exercise, giving the story of the Bible, testing knowledge of content and relationships.

Administration: Age of group: Form I Grades above III, Form II high school and adult.

Time required: About 50 minutes.

Special conditions: No time limit.

Scoring: Tentative scores based on answers of theological students at Union Theological Seminary, The University of Chicago Divinity School, and on results from 300 pupils. Point of view of modern scholarship assumed. Score sheets furnished.

Standardization:

Reliability: 0.90 (Form II).

Validity: Correlation of 0.25 with intelligence.

Norms: Being gathered.

Other Comment: For later developments consult:
Goodwin B. Watson,
Teachers College,
Columbia University,
New York, N. Y.

Name of Test: Upton Chassell Citizenship Scale.
Publisher: Bureau of Publications, Teachers College, Columbia University, New York, N. Y.
Price: 2 cents for short forms in groups of 25 or more.
Purpose: To provide a scale upon which public school teachers can record the social development of pupils.
Contents: Takes seat in assembly quietly.
Avoids getting wet, getting chilled, or cooling off too suddenly after play.
Does not worry.
Verifies tentative conclusions on the basis of further observation and experimentation.
Follows the rules of the game scrupulously.
Appreciates other nations and races and their contributions.
Administration: Age of group: Teacher can apply to any pupils.
Time required:
Special conditions:
Scoring: Gives score in points and in per cent. The various short scales are comparable.
Standardization:
Reliability: 0.81 to 0.97. Correlation of teacher rating with self-rating is 0.50.
Validity:
Norms:
Other Comment: For full directions, see *Teachers College Record*, Vol. 23, pp. 52–79, January, 1922.
For later developments consult:
Dr. Clara Chassel Cooper,
227 Summit Ave.,
Richmond, Ky.

Name of Test: Woodworth-Mathews Personal Data Sheet.
Publisher: C. H. Stoelting and Company, 424 N. Homan Ave., Chicago, Ill.
Price: Woodworth (for adults) 9 cents each in orders of 25; Mathews (for adolescents and children) 5 cents each.

Purpose: To show the general emotionality, nervous, and mental stability of adolescents.

Contents: 100 questions such as:

Do you often sleep poorly? *Yes. No.*

Are you ever troubled with dreams of people being dead, or about robbers? *Yes. No.*

Do you worry too much about little things? *Yes. No.*

Does it make you uneasy to cross a bridge over a river? *Yes. No.*

Do you pick your nose? *Yes. No.*

Do you find it difficult or impossible to make love? *Yes. No.*

Administration: Age of group: Woodworth: Adult. Mathews: Adolescent, pre-adolescent.

Time required: About 50 minutes.

Special conditions: Should be used only by persons competent to handle individual emotional problems in a scientific way.

Scoring: Total number of significant answers.

Standardization:

Reliability: 0.55 to 0.90 depending on investigation.

Validity: 0.12 to 0.66 with criteria of emotionality based on ratings.

Norms: Available at extra cost of 25 cents each for Mathews Tests.

Other Comment: See Franz's "Handbook of Mental Examination Methods."

Also, *Journal of Delinquency*, Vol. III, No. 1.

For later developments consult:

Professor R. S. Woodworth,
Columbia University,
New York, N. Y.

Inasmuch as only tests in published form were listed in this first discussion, there are, of course, many other valuable tests which should be mentioned.[1] It is particularly worth while to know certain suggestive features of tests which are now no longer available. Makers of new tests may wish to refer to some of the items contained in the early and ingenious developments. There are also certain other tests in process of construction in which are embodied good suggestions for test makers. This

[1] On June 1, 1926, C. H. Stoelting and Company, 424 N. Homan Ave., Chicago, Ill., announced the following tests in preparation:

TJADEN, J. C., Analytical Interview for the Study of Individual Delinquency.

DAVIS, F. K., Anxiety Test.

JONES, E. S., Personnel Questionnaire.

OTIS, MARGARET, Suggestibility Test.

second group of tests is listed in alphabetical order under the name of the institution or person responsible for its development.

4. Tests Containing Valuable Ideas, but Not Now Available.

a. Case's True-false Test in Religious Education

This test of 100 true-false elements forms the basis for Professor Case's book "Liberal Christianity and Religious Education."[1] Some of the items deal with a liberal interpretation of the scriptures, others with liberal theology, and still others with a liberal social and economic point of view.

b. Chapman's Test of Motives

Chapman presents to pupils a group of 10 reasons each, for going to high school, for saving money, and for reading good literature. These reasons were developed by study of compositions which pupils wrote on these themes. The test consists in asking pupils to arrange them in order of their strength. The best order was determined by the judgment of competent adults. With this test he has found a steady progress through the grades in ability to arrange the answers in an order which correlates highly with the criterion.

c. Character Education Inquiry

The Character Education Inquiry at Teachers College, Columbia University, under the direction of Hartshorne and May, has developed the best series of ethical tests which has yet been attempted.[2] A test of "opposites" having a reliability, if lengthened to require one hour of time for taking, of 0.96 was discarded because of the Schwesinger study in social-ethical vocabulary. A "similarities" test was likewise developed but not used. The "word consequences" test asks subjects for the most probable consequences of such acts as lying, gambling, etc., and for the selection of best and worst among those consequences. A "cause and effect" test studies perception of the relationships of social phenomena by a true-false test having a reliability for one

[1] Case, "Liberal Christianity and Religious Education," The Macmillan Co., New York, 1925.

[2] They published some material upon their tests of right and wrong in *Religious Education* during 1926 and 1927. This is now published as a complete monograph by the R. E. A., 308 N. Michigan Ave., Chicago, Ill.

hour of 0.88 and a correlation with unweighted criteria of 0.52. The "duties" test asks subjects to discriminate among acts which are or are not moral duties, this test having a reliability of 0.95 for an hour's time, and a correlation of 0.54 with the criterion. The "comprehensions" test asks pupils which of several alternatives they would choose in certain typical situations, and has for an hour, a reliability of 0.90 but a correlation of only 0.37 with the criterion. The "provocations" test endeavors to stretch conventional moral responses by introducing counter-provocations to see what the subject believes it best to do in such a situation of conflict. The reliability here would be 0.90 for an hour, the correlation with criteria 0.42. The "foresights" test endeavors to obtain pupils' notion of consequences when none are suggested; the "recognitions" test measures ability to classify correctly certain acts under general terms like dishonesty or impurity (reliability 0.92, validity 0.58). The "principles" test studies knowledge of conventional ethical principles by a true-false method (reliability 0.92, validity 0.64), and the "applications" test (reliability 0.91, validity 0.42) measures ability to apply these principles to situations. Out of this total battery significant test elements have been assembled in new and shorter forms, called "Moral Knowledge Tests" and are undergoing further study. Among the interesting results so far published have been findings showing little correlation between such moral knowledge scores and the conduct scores obtained in tests of deception and of helpfulness. Correlations between total score of pupils and of others who might influence them, show that pupils tend to resemble in size of moral knowledge score (as in intelligence) most of all their parents, next their friends, and little if at all, their school teachers, Sunday school teachers, or club leaders.

d. Chassell's Parable Interpretation Test

This was printed in *Religious Education*, for December, 1921. It suggests that the various parables of Jesus be read to pupils. After each reading the pupils check the best among several suggested meanings for the parable. The alternatives are easy and suitable for children who can read. They have occasionally been

used with children too young to read who have taken the test orally.

e. Fahs' Biblical Test

This test, still in the experimental stage, seems to be one of the most complete and suggestive tests of Biblical understanding which has yet been developed. Among the tasks which it sets before the subject are:

1. To classify certain names, indicating whether the name is that of a city, country, mountain, man, woman, lake, or book.
2. To fill in blanks in statements about Bible characters, from which the names have been omitted.
3. To arrange in chronological order the names of Bible characters in several groups of five each, this testing understanding of the general course and sequence of Hebrew history.
4. To check the statement which seems to the pupil more important in each of the number of pairs of statements, one of the statements representing a fundamentalist point of view, and one a liberal point of view.
5. To indicate after each of a number of Bible incidents whether the person described in the incident did right or wrong.
6. To check after each of several Bible phrases, a statement of meaning among several possible ones, which most truly agrees with the meaning of the phrase.

f. Fernald

Among many other tests, Fernald has developed a list of meritorious acts, 10 in number, and a list of ambitions, each of which can be arranged by a subject in order of preference. The correct order is determined by a series of judgments by court judges and other persons supposedly competent. A person who can arrange these acts or ambitions in the order approved makes the highest score.

g. Hartshorne in 1919, and Miss Tracy in 1924, experimented with the unfinished story as a test of moral character. With first-grade children Miss Tracy told a story of a boy whose mother asked him to go to bed quietly and quickly by himself,

then left him all alone. She asked the children individually to finish the story for her, and found that the replies were significant in many cases of the viewpoint of the child. Some answered, of course, that the child got up, played with his toys, while others thought that the boy was afraid to be left alone. It was found very difficult to distinguish with these tests between the children who tended to report what they would do, and the children whose love of a good story led them to imagine a series of wild offenses which they would never perpetrate.

h. Self-ordinary-ideal Rating Scale

This has been used by Knight, Franzen, Kinder and other persons and is at present published as Part II of the Emotional History Record summarized previously. The subject is asked to rate the interest which he, himself, has in each of a number of lines of activity, and also to rate the average person, and the ideal person on the same interests. The agreement between the rating which any individual gives to the average person in the group and the real average as obtained by averaging self-ratings is an index of insight. The amount of difference between an individual's self-rating and his ideal ratings seem to be a good measure of conflict. The extent to which any individual's pattern of self-interests agrees with the average of his group is a good measure of conformity. The disagreement between an individual's ideal and the average ideal of the group in which he lives seems to be a good measure of difficulty of adjustment. If the deviation between self and ideal be subtracted from the deviation between average and ideal, the difference appears to be a measure of self-esteem. If it is positive an individual must think himself nearer his ideal than is the ordinary person, whereas if it is negative, he feels the average person to be more nearly ideal than he is. The author, in cooperation with Mr. Chassell, is carrying on further work in standardizing this rather widely useful test.

i. Laslett's Controlled Association Test

Raubenheimer tried out upon delinquent boys and non-delinquents of similar intelligence a number of words such as "Teachers." Some boys answered "They work hard," others, "They know they can punish you," others, "They are kind of

cranky," still others, "They are not fair to you." Laslett experimented further with such a list and selected words on which there was a maximum difference between the delinquents and non-delinquents. He obtained a reliability of 0.77, and a correlation with ranking of 0.39. For instance, when he said "bar" some people thought "candy" and some thought "saloon," or "jail."

j. Orr's Good Manners Test

Miss Orr has only recently developed her "Good Manners Test." It is reported in connection with the Character Education Inquiry articles in Religious Education. It is built in multiple choice form, and suggests several actions for each situation. The pupil is given an opportunity to choose the best among these. It seems to be a fairly good index of home background and training.

k. Porter's Advanced Bible Test

This test is still in process of study by Porter, but it bids fair to furnish the best advanced examination in Bible material which has been developed. The following questions are suggestive:

Directions: In each of the following questions find the one best answer among those suggested. Draw a circle around its number.

2. The international horizon grew suddenly greater and more terrifying to the people of Jerusalem in the days of
   (1) David. (2) Noah. (3) Ezra. (4) Isaiah. (5) Jacob.
6. Worship of a more spiritual character than that of sacrifice appeared as a result of
   (1) the division of the kingdom; (3) the conquests of Alexander;
   (2) the Babylonian exile; (4) the conquests of Pompey;
   (5) developing city life in Jerusalem.
4. The home rule allowed the Jews by the Romans in the days of Jesus was controlled by the
   (1) Sadducees. (2) Pharisees. (3) Samaritans. (4) Essenes.
   (5) Publicans.
8. The situation which existed in Judea and Galilee in the days of Jesus resembled which one of the following modern ones?
   (1) American readjustment of religion to scientific discoveries.
   (2) The rise of nationalism and opposition to British control in India.
   (3) Drift away from the church in modern France.
   (4) Reaction against socialism in Italy under Mussolini.
   (5) Chinese progress in democracy.

l. Porter's Student Opinion on War

This is a compilation of statements about war, its causes, consequences, and cure. Some of them are extremely militaristic in viewpoint, others extremely pacific, while others fall somewhere between. The subject is given an opportunity to state his attitude toward each one upon a five-point scale. Porter has selected, on the basis of the replies of individuals of known points of view, those items which are most valuable in clearly differentiating the prevailing points of view. Other tests of attitude upon international questions have been developed by Manry (University of Iowa), Neumann (Teachers College, Columbia University), Keeny (*The Inquiry*), and Ewing (National Council of Y M C A).

m. Raubenheimer's An Experimental Study of Some Behavior Traits of the Potentially Delinquent Boy

Psychological Monographs No. 159, 1925, Psychological Review Company, Princeton, N. J. Most of the tests presented by Raubenheimer in this excellent study are conduct tests and will be discussed later. Two or three seem to be primarily pencil-and-paper tests. One of them is a character preference test. A certain type of individual is described in concise and simple manner: "Buck is 15 years old and is a real dare-devil. He does not care much for school work and wants to get out. He often goes off with the circus, and gets around by picking up rides. He thinks he will get away right now if he can only find a chum to go with him." The subjects are asked to rank these boys in the order in which they would prefer them as chums. The test seems to have a reliability from 0.64 to 0.80, and correlation with rankings of about 0.18. Raubenheimer has used, as did Cady, Fernald, and others, an offense rating scheme in which a series of brief descriptions of bad conduct are rated in order of seriousness. These seem to be about as useful as the ratings of meritorious acts described above.

n. Schwesinger's Study of Socio-ethical Vocabulary

Miss Schwesinger selected from the Thorndike and other word lists, one thousand words which had social or ethical connotations. These were tried out on a number of children, and their

order of difficulty determined. They have been built into a series of multiple choice, definitions, but the exact difficulty of each has not been determined. It is known that the test is an excellent measure of intelligence and correlates fairly well with other measures of home background. It is also reported in the Character Education Inquiry work.

o. Travis' Study of Personality Traits

Travis selected from a large number of lists of significant traits of personality 25, and made two statements briefly describing each. He then grouped the 50 descriptive phrases in five groups of ten each. He asked subjects to arrange each group of ten in order of their preference. He found a very high relationship between the order in which one series of statements for the 25 traits was arranged, and the order in which the second series of statements representing the same 25 traits was arranged. He found that the order gave a correlation of about 0.30 with rankings.

p. Van Wagenen's Tests of American History, Character Judgment Scale

In a scale, now out of print, Van Wagenen employed an interesting device to discover the kind of motives which children would project upon other persons. He described an historical incident, say, for example, that of a soldier who crept out of the lines into the enemy territory, and returned bringing several captives single handed. He then asked the children which of several reasons represented their notion as to why this soldier did this feat. Was it perhaps in order to show that he was not a coward? Because he thought the general would praise him? Because he had been angered by some of the enemy, and wanted to get even with them? It seems quite probable that the motives which people will assign to other people will, in some degree, give away the motives which they, themselves, would be most likely to feel.

q. Woodrow has described[1] a "picture-preference test" for children. He uses sketches of children and adults doing various desirable or undesirable things. In one picture a child may be

[1] *Jour. Ed. Psych.*, Vol. XVII, p. 519, November, 1926.

helping with the dishes, in another he may be catching a ride on the back spare tire of an automobile. Such pictures are presented in groups of four and children are asked to pick out those liked best. No criteria, moral or otherwise, are offered. He found the test to be reliable for the third grade, to the extent indicated by a self-correlation of 0.79. The relationship to intelligence was 0.24 and to ratings on character 0.41.

r. Y M C A Religious Education and Character Growth Tests

Probably the most complete work in testing of character elements which has been done by any religious organization has been carried forward by the Y M C A as a development out of the old Bible study examinations.

Beginning in the winter of 1923–1924 these examinations were put upon an objectively scorable basis, and made to include measures of religious ideas and of ethical judgment. Many of them were in the form of life situations in which the boys were asked to decide what they believed to be the best thing to do. These have been given to 5,000 or 10,000 boys a year, and more than 100 different examinations have been constructed. Unfortunately, sample copies of the work are not available, but a publication[1] of the National Council of the Y M C A, 347 Madison Ave., New York, N. Y., describes in some detail some of the better questions. Among the more interesting test devices which have been used is one which asks the individual to arrange the elements of an incident in the order in which they probably happened. Thus, for example, boys are asked to tell which of these happened first, which second, which third, and which last.

Parents of certain "better" families forbid their children to play with Negroes.
Negroes play a basketball game with the white boys.
White boys invite the Negroes to a club party.
Parents give their consent for Negroes and white boys to play together.
Negroes move into the neighborhood.

It is fairly clear that the arrangement of these elements in certain orders represents more insight into the way in which race

[1] "Experiments with Religious Education Tests," Program Papers No. 5, Association Press, New York, 1926.

relations, race feelings grow up and change, than would arrangement of the same elements in other orders.

Another interesting and somewhat different test was developed as a measure of the results which boys in the Y M C A obtained from studying a book entitled "The Spirit of World Brotherhood." This book calls for discussion of various questions but does not offer any dogmatic solution. The test, therefore, presented a large number of points of view with reference to immigration, world court, racial equality, international treaties, self-determination, etc. Each statement was paralleled by another equally plausible, but which presented an incompatible point of view. Thus, if one believes that, "The best interests of every race require that it preserve its own purity of blood, and not intermingle with other races," then one can hardly at the same time take the position that, "The most-to-be-desired goal of mankind is a mingling of all races in one blood which shall unite the contributions of each, eliminating, so far as possible, the defects of any, and doing away with the racial differences and distinctions." These statements were all jumbled together throughout the test and an individual was credited for each consistency, and debited for each inconsistency. Thus, no particular point of view was measured, but only an insight into the real differences between thinkers, and the ability not to try to stand on both sides at once. Still another interesting test found among these Y M C A questions is one which gave a series of Bible incidents, and asked for a judgment as to the rightness or wrongness of each. Each one of these incidents was paralleled by another modern life incident in which exactly the same principle seemed to be involved. The modern life incidents and the Bible incidents were all mixed up together in the text, and again pupils were credited for consistency in approving or condemning in modern life the same sort of thing which they approved or condemned when it was done by a patriarch of old.

Beginning in the fall of 1926 the Y M C A is promoting a nation-wide scheme of character growth tests. These tests consist of one form to be given to boys and young men in the fall, and a second parallel form which is to be given in May or June

to measure changes which have taken place during the year. Tests have been prepared for three age levels, boys 12 to 14, older boys 15 to 18, and young men over 18. At each age level the tests cover five aspects of character. One test measures self-esteem, adjustment to group, insight into others, etc., as suggested in the Self-ordinary-ideal Rating Scale (p. 100). A second deals with immediate personal relationships in home, church, school, and vocation. A third tests attitudes in relation to the club group and its activities. A fourth deals with religious information and concepts, and the fifth with Christian World Citizenship, including international, interracial, and industrial problems. The tests prepared for 1927–28 deal with ethical discrimination, sense of values, emotional stability, and creative power. These tests will be available to groups outside the Y M C A where desired.[1]

5. Life Situation, Conduct, and Behavior Tests[2]

A third general group of interesting tests contains those tests which are primarily tests of behavior and conduct. In a sense, all tests are a measure of behavior, sometimes with vocal muscles, sometimes with paper and pencil, sometimes in more varied fashion. There is a tendency, however, for the tests which are now published in printed form to deal very largely with ideas as contrasted with actual conduct. The Downey Will-temperament tests are an exception to the ordinary practice. The group of tests presented below deal much more definitely with the way in which subjects behave in certain special situations. From these it may be possible to make some inference as to the reasons for their behavior, and as to how they will behave in other situations. These tests have been listed under the names of the traits they are supposed to measure. This is a rather dangerous procedure. It suggests that the test may really measure the thing for which it is named. It is necessary to exercise great caution in making this assumption. The tests measure behavior in specific and limited

[1] Address inquiries to A. J. Gregg, Chairman Committee on Tests, National Council of the Y M C A, 347 Madison Ave., New York, N. Y.

[2] References to original articles discussing each of these and other tests may be found in bibliographies previously published, and will not be detailed here. See note on p. 71.

situations. How far that represents a general characteristic depends, in large degree, upon the number and variety of these situations.

a. Aggressiveness

Tests for aggressiveness developed by Moore measure the length of time which a person will spend at a task when certain distractions, such as the constant expectation of an electric shock, or a snake 9 inches from the face (later eliminated), or a command to look the investigator squarely in the eye, have been introduced. He found that persons selected as the most aggressive in the class were much less upset over such things, than were the least aggressive. He found that the number of times the subjects shifted their gaze away, while they were supposed to be facing the examiner and performing simple mental arithmetic, was an excellent index. Combining all of these tendencies he stated that he had a better measure of aggressiveness than the army alpha offered for intelligence.

b. Caution

Caution has been measured in terms of the ratio between the number of attempts and the number of errors on a very difficult test in which subjects have been told they will be penalized for guessing.

c. Cheerfulness

Miss Washburn and others have measured a tendency toward cheerfulness as opposed to depression by methods of memory and free association. Subjects may be asked to recall as many emotional experiences as possible without regard to the type of emotion. There seems to be a correlation between the number of pleasant experiences reported and ratings which were given to people on cheerfulness. Probably the better method is the one which propounds to the subject a list of neutral words such as sky, color, chair, man, feel, etc. Subjects are asked to give the first word or phrase which comes to mind, and if that is not a distinctly pleasant or unpleasant association, to continue to think of associations until one is reached that has a distinctly pleasant or unpleasant tone. It has been found that people who are rated by their associates as generally cheerful, are much more

apt to think of words having a pleasant-feeling tone. Sixty-five per cent of the persons who stood highest in ratings for cheerfulness also stood highest in the number of pleasant associations when responding to a list of 50 words.

d. Civic Duty

An unknown person developed this test with a group of boys he took for a day of sport to a public park. Several of the games involved the use of slips of paper upon which each subject's name was written. Nothing was said about picking up the paper, but a waste-can was conveniently located. After the games were over and the boys had gone home, an assistant went over the ground and picked up all the pieces of paper which had been thrown on the ground, each paper bearing the name of the boy who had left it there.

e. Concentration

Other investigators have measured a quality called concentration and persistence. They have placed children in a room where there were a great many interesting and distracting books, pictures, and articles. These children have been instructed to carry on a monotonous task, and are left to themselves. The amount which the children accomplish under such circumstances compared with what they can accomplish under definite adult supervision forms an index of their ability to resist distraction and to concentrate effectively on their work. At least one attempt has been made to make a test like this in printed form, using a central column in which letters are to be crossed out while the borders of the page are filled with jokes, cartoons, and interesting distractions which are supposed to be ignored. Voelker had his subjects count *A*'s during five minutes in uninteresting printed material, and then count *A*'s in similar printed material in an illustrated picture book. He tested the dependability of a subject further by leaving him in a room with instructions to push a certain button every 2 minutes. There were interesting objects in the room. An assistant at the far end of the wire, in a remote place so that the boy could not know that the signals were recorded, took down a record of any omissions or delays in this task of pushing the button.

f. Confidence

Self-confidence has been tested largely in terms of the certainty which one has that one's judgments are correct. Trow used lines, fallacies, and ethical situations. In each case he asked the subject to indicate how certain he was that the judgment made was correct. Whipple employed a similar principle in his early tests of fidelity of report. He showed pictures to a group and then asked them to report what they had seen in the picture. He gave them a chance to underline and double underline statements about which they felt unusually certain. He also asked a series of questions, some of which pertained to objects in the picture, others which were misleading, and had these answers rated for certainty also. He found wide individual differences not only in memory capacities, but also in willingness to assert with confidence.

g. Conformity

Deutsch has developed some of the best tests for conformity. Subjects are asked to state their preference, among certain drawings, certain types of beauty, marriage customs, transportation systems, popular superstitions, ideas about immortality, etc. In each case there is one which is very common in this particular civilization and generation, others which belong to remote times or other races. Conformity is measured by the number of selections which are true to the prevailing type within the group. Non-conformity, on the other hand, is evidenced by the choice of unusual and exotic alternatives.

h. Courtesy

Mrs. Bonser developed with children about nine years of age a test of courtesy in a social situation. The children planned a party to which their parents were to be invited. Previous to the party the pupils took a pencil and paper test in which they indicated their judgment as to what a child ought to do in a party situation. Unknown to the children some of the guests were chosen as competent observers and asked to make a careful report of just what the children did by way of greeting strangers, offering food, giving up seats, helping to entertain, showing-off, and other forms of desirable or undesirable action. In the small

group tested there seemed to be comparatively little relationship between what the children believed it good to do and what they actually did. Further experiments in party situations have been made by the Character Education Inquiry. Voelker used as a test of trustworthiness a test which in many ways seems to be better classified under courtesy, in which he sent boys on an errand and offered them a tip. Those boys were credited who refused to accept it, while partial credit was given to those who protested but finally took it.

i. Decision

Bridges has studied speed and accuracy of decision by presenting the subject with an apparatus which presents at a given signal two cards. On one of these may be a picture of an apple, on another a picture of an orange. The subject is to decide which he would rather have at the moment. The time taken to make the decision can be measured with mechanical accuracy. Not only fruit, but also colors, designs, faces, and ideas represented by such words as money, leisure, prestige, excitement, etc., were used. Accuracy of decision is measured by asking a subject which figure is larger, which card has more holes punched in it, and similar questions in which the decision can be checked against actual fact.

j. Delinquency

Many of the tests listed below under honesty and suggestibility have been applied to delinquent and non-delinquent children in the endeavor to find tests which will clearly separate one group from the other. One of the best attempts to discover and develop character tests which will differentiate delinquent groups was made by Lentz. He tried out Pressey tests, Koh's ethical tests, and numbers of others upon two groups, one of which was delinquent, the other of which had the same intelligence and the same home background, but which was not delinquent. Upon this basis he selected certain tests, but when it came to applying the tests upon two other groups, he found the basis that worked the first time did not necessarily work the second time.

As a result of this check work, only two of the six tests were shown to indicate valid and significant differences between the delinquent and non-delinquent groups.

These two tests were a short questionnaire and what is termed a "Daily Contribution" test. The questionnaire covered such points as church attendance, number of musical instruments and rooms in the home, working for pay, preference for school or work, etc. The delinquents invariably express a greater preference for employment over school attendance; likewise, report less church attendance, and fewer musical instruments and smaller homes. The "Daily Contribution" test called upon the boys for a daily response in supplying the examiner with information in a general enterprise for collecting certain data. As judged by this test there seems to be a striking difference in the cooperativeness of the delinquent and unselected boy. While the test as used at times failed to function and obviously is a very rough test in need of much refinement, it gives promise of being an excellent lead in this field.

A significant feature of the experiment was the failure to find any significant difference between the delinquent and non-delinquent in tests of ethical discrimination. If this finding repeats itself in further studies it will relegate the Ethical Discrimination test to the rank of a poor test of intelligence.

There are other interesting indications from the preliminary study which are not so final since they were not checked by the larger subsequent groups. Four honesty tests were administered in which the percentages of "dishonest" reactions were consistently lower among the delinquent group. Further verification of this point might prove an enlightening commentary on modern education. A repetition test was given in which the situation calling for the same reaction was presented twice and the discrepancy between the two reactions noted. The situation included several lists of words in each of which the subjects were asked to cross out those things in which they were interested.[1] According to the results of this test one would be inclined to say that the delinquents were more inclined to change their mind, to feign an interest oftener, or to have fewer real interests. The author believes this technique may eventually lead to the successful testing of interest in various fields. A list of occupations was submitted for rating for general social usefulness. It was interesting to note that both delinquent and non-delinquent groups rated doctor first and movie actor lowest in usefulness. This same list was later presented for rating for preference as a personal vocation. The discrepancy between these two ratings seemed to show some difference between the groups. It was thought that this test might have some bearing on vocational altruism. Other tests reported but not checked by the final groups should prove stimulating to other workers in the field.[2]

[1] Adapted from Pressey X–O Test. See p. 89.

[2] Lentz, "An Experimental Method for the Discovery and Development of Character Tests," Teachers College, Columbia University, New York, 1925.

Raubenheimer, in his study mentioned above, tried out a number of conduct tests as well as paper and pencil tests, finding that it was possible to differentiate delinquents from non-delinquents with very excellent success, except perhaps in the case of delinquent boys from high-class homes.

k. Emotionality

The most successful tests of emotionality so far developed make use of the fact that in state of emotion the heart beat tends to be quickened, blood pressure rises, breathing is more rapid, and the electrical conductivity of the body is increased. All of these changes can be measured with proper instruments. Careful experimentation has shown that it is possible to recognize strong emotions almost without fail. It is even possible to note changes in the case of such subtle emotions as those which accompany the consciousness of having told a lie. It must be remembered that it is the emotion that is measured and not the extent of truth or falsity. Some people might be very much upset over a minor distortion of the truth whereas others would feel comparatively little affective change in situations of much greater gravity. Another measure of emotionality was mentioned under the heading of cheerfulness. The number of emotional experiences which a person can recall within a given space of time seems to correlate well with ratings on emotionality. Another persistent investigator followed people in social situations and with a stopwatch recorded the amount of time that each one spent laughing. This proved to be the best of several measures he obtained on general emotionality. Another type of test measures particular emotional susceptibility as well as its general level. This was developed by the author, and consists of prose paragraphs alike in the average number of nouns per line, alike in general reading and comprehension difficulty, alike in length and general form, but very different in content. One paragraph dealt with ordinary material which would be non-affective, a second page was a sentimental description, a third was humorous, a fourth was foreboding, gloomy, and fearful, a fifth was vehemently anti-religious, a sixth suggested passionate love, a seventh intense grief. Subjects were asked to cross out in each paragraph as many nouns

as possible within a period of two minutes' time. Marked deviations from the general curve of improvement can usually be attributed to interference with the business of crossing out nouns by the emotional state aroused by the material. This was particularly useful with a group in which the material could be rotated so that practice effect was equated for the group as a whole, and the marked deviations which still ensued seemed to be safely attributable to the emotional state which the material produced.

l. Group Loyalty

McCaskill, working with the Survey of the Y M C A in New York City developed a test for group loyalty which is very ingenious. He told the boys that they were going to have a series of tests on their knowledge about sports, about baseball, basketball, football, and the movies. He announced that a prize would be given to the individuals making the highest scores, and also to the club groups making the highest scores. He gave each boy a sheet on which it was indicated that form A counted 10 points for the individual and 90 points for the group, form B counted 20 points for the individual and 80 points for the group, Form C counted 40 points for the individual and 60 for the group, form D 50-50, form E 60-40, form F 70-30, form G 80-20, and form H 90 for the individual and only 10 for the group. He told the boys that there would be no difference in the general content of the tests they might choose. Each boy was to check three forms which he wished to take. These slips were then turned in and the boys were given, in accord with their request, the three forms of the tests. The sports knowledge test proved an interesting diversion, but the real measure of group loyalty he believed to reside in the tendency of some groups to choose mainly tests that would work toward a group prize while other groups in the same situation chose only tests that counted mainly for the individual.

m. Helpful Behavior

The Character Education Inquiry has utilized several tests of willingness to help others at some cost to oneself. These tests sometimes require giving up of a certain part of the money

available to the individual in order to help children in the Near East. Sometimes it is a matter of going without ice cream for a similar cause. On one occasion neat boxes containing pen, pencil, eraser, etc. were presented a class, and soon after, appeals made for sharing with children who had none.

n. Home Environment

Most studies of home environment have been based upon rating scales. A number of rating scales are in existence of which perhaps the best known is the Whittier scale for grading homes, and the Whittier scale for grading neighborhoods. Both are obtainable at the Whittier State School, Whittier, Calif. Sims has recently developed a scale for home environment which is much more objective. He has found statistically the importance of each of a number of elements in determining social-economic status. He found the most important ones to be: having a mother who had graduated from high school; having 125 books or more; having a piano; and subscribing regularly to two or more magazines. Each of the many elements in the scale is rated in accordance with its significance as an indication of the general status of the home. The scale has a reliability of 0.77. The Character Education Inquiry are developing another test of home status which is somewhat more indirect and deals with the kind of information and attitudes which a child can obtain in his home environment.

o. Honesty

Under this general heading are included also the many tests of trustworthiness, reliability, tendency to exaggerate, to cheat, etc. Many of the tests reported here were developed by Voelker, used further by Cady, then by Raubenheimer, later by Ruch, Terman, the Character Education Inquiry, and other investigators. To Voelker in his dissertation on "The Function of Ideals and Attitudes in Social Education"[1] belongs the primary credit for originating conduct tests of these traits. The best and most extensive work upon such tests has been done by Hartshorne and May of the Character Education Inquiry.

1. The overstatement test. Pupils are given certain marks

[1] See footnote, p. 52.

upon their school work. Later in a test or an interview they are asked: "Did you receive 95 per cent in arithmetic in your last test?" the grade being higher than the one the subject really received. Raubenheimer obtained a reliability of 0.56 which he was able to raise to 0.86 by revising this test. Correlation with ratings was 0.50.

2. The subject is given certain work to do, a puzzle to solve perhaps, is promised a certain credit for it, and agrees to solve it without help. In a later situation a different examiner offer to give him help. The subject is credited only if he refuses.

3. Subject is given an opportunity to take some little article which he would like to have. Thus, he is left with a box of fascinating puzzles which he could slip into his pocket if he so desired. This is later checked up. He is sent on an errand down the hall, and a pocketbook is left lying near his path. The pocketbook has a little small change in it. This, too, is later checked up.

4. Subject is sent to borrow some material, but before he is given it he agrees to return it by nine o'clock the following morning. The test is passed successfully if he returns the material as promised.

5. Subject is sent on an errand to a store, and is given ten cents over change. He is credited only if he returns the change. If he turns over all the change to the examiner without noticing that it is ten cents too much, the examiner calls his attention to the fact and asks: "Did you leave some of your own change in with this?" to which, of course, the boy is supposed to answer "No, that is what the man at the store gave me."

6. Subject is asked to do certain tasks with his eyes shut. It may be to put together a certain puzzle simply by feeling (profile test) or to put crosses in the center of some small squares and circles scattered over a sheet of paper. Each of these tasks is impossible with the eyes closed so that if a subject succeeds, it is certain that he peeped.

7. Subject is given an examination on a four-page folder. The back of the second sheet is covered with a paraffin solution so

that all the marks which the subject has put down are recorded on it. This waxed page is torn off and turned in as a separate part of the test. Then subjects are given a key to the right answers and allowed to correct their own papers. The difference between the score recorded on the waxed paper which was what they originally did, and the score on the paper as they turned it in is a measure of the extent to which they have been incorrect or dishonest in their work.

8. Cheating is also frequently tested by giving pupils a certain test with an opportunity to score themselves upon it, then in some measure checking up on the score which they gave themselves. Sometimes the original papers are collected and gone over again; sometimes the subjects are given a retest on the same material, the material being of such a simple nature that any wide difference between a person's score at one time and a score under more careful supervised circumstances is very suspicious.

9. Subjects are given a chance to overstate their knowledge or ability. Perhaps the most frequent form of this test is the conduct test. People are asked to check the books they have read in a certain list. Half of the titles are real, but half of them are fictitious and do not even resemble the titles of existing books. With a certain small allowance for error, the number of books checked in the non-existent titles is an index of the unreliability. Again subjects may be asked to answer such questions as "Can you swim?" "Can you typewrite?" "Do you know the capital of each state?" "Do you know the names of the continents?" "Do you know the number of yards in a rod?" Then the hard-hearted examiner gives out a card which requires that the subjects give the capital of each state, the names of the continents, and the number of yards in a rod. Opportunity is also given for subjects to show their ability to swim and to typewrite. Recent investigation suggests that this test is an excellent measure of "general character."

10. Pupil is given a certain errand to carry out, say the delivery of an envelope containing a small amount of money, or a question about the boy's conduct at home. He is asked if he will deliver it without reading it. If he agrees to do so, and then in some

manner tampers with the contents or fails to deliver it, he is scored for untrustworthiness.

11. A form of test which occasionally finds publicity in the Sunday paper is that employed by the man who sent a dollar bill to each of twenty friends with a note saying he was sorry to be delinquent in paying back what he owed. In at least one experiment the man reported that not a single person returned the dollar bill although he owed it to none of them. Voelker modified this and improved it, by sending 25 cents through the mail to boys thanking them for helping in the boys' advertising campaign. He was careful, however, to misspell the name and misstate the address so that careful inspection would show that the letter was not originally intended for the person who received it. There was a provision made for a return acknowledgment and he scored as somewhat dishonest, boys who kept the money.

12. Subject is playing a parlor game in which the race rules require carrying one bean at a time. In order to win, some pupils will try to carry several. This can be modified to fit various competitive games.

13. In an athletic contest, pupils are allowed to report for themselves certain scores for strength. Previous or subsequent tests by the examiner reveal gross exaggerations.

p. Honest Confession

This is a rather ill-named trait, but it embodies that too familiar tendency of Sunday-school pupils to give Sunday-school answers. Thus, the Y M C A, out of 2 years of experimentation, found that the best single question among several hundred, for differentiating boys who made high scores on the whole, from boys who made low scores on the whole, was this true-false statement, "A boy should always do whatever his parents tell him to." Boys who answered "No" to that statement averaged higher in general score than did the boys making any other answer in a similar true-false question, It proved, therefore, desirable to compile a test which embodied a great many such statements. Mark May of the Character Education Inquiry originated the further suggestion that a test should embody things which people ordinarily think of themselves as expected

to do, but which nobody really does. Thus, boys were asked whether they picked up every paper which they saw in the street, whether they ever kept a pencil that belonged to anybody else, whether they invariably asked all their friends to come to Sunday school, etc. Each of these was stated in so extreme a form that a truthful answer must necessarily be a confession. These questions scattered through the tests which embody many other types of question can then be picked out and considered as a unit, and give a fairly good measure of an individual's tendency to answer what he thinks he is expected to answer rather than to answer truthfully.

q. Humor

One of the common tests of humor is the practical joke played on an individual, the assumption being that if he has a good sense of humor he will laugh the matter off. Within limits this is probably a fair test. Another form of humor test that has been developed makes use of the Healy Pictorial-completion Test I (C. H. Stoelting and Company). The ordinary instruction is to put in the most appropriate object for each vacant space. As a test of humor, however, the instruction is to put in the funniest object at each point. A little preliminary experience has shown that the subjects which children consider to be funniest change very distinctly with the child's age. A third form of humor test presents the first part of a joke with several possible endings, one of which is presumably much funnier than any of the others. Subjects are supposed to check the one which would make the funniest joke. The difficulty is to eliminate the influence of the number of jokes heard. It is very difficult to get situations of this sort which do not represent the common experience of readers of magazines and newspaper columns.

r. Imagination

One of the best tests of originality and imagination has the unfortunate qualification that the scoring scheme has not been standardized. Subjects may be asked to give words as different as possible from such words as "heaven," "die," "Russia," and "hard-boiled egg." The direct opposite is obviously the poorest answer, but there are certainly degrees of "best" which it is

difficult to define. McGeoch found that imagination was very well measured in terms of the number of associations which people could make with ink blots of odd shape, and the number of words which could be built in limited time out of letters contained in two six-letter words.

s. Interest

Several measures of interest have been developed, aside from the type of direct-interest questionnaire discussed previously in connection with the interest analysis used in personnel research (p. 84). The Y M C A, in connection with their summer-camp examinations, modified a scheme used by Raubenheimer, presenting brief case descriptions of boys, one of whom was interested primarily in stars, another who liked nothing better than swimming, a third who especially enjoyed getting dates with classy girls, a fourth who spent most of his time at his school work, etc. The subjects were asked to rank those boys in the order in which they would prefer them as chums. It was felt that since the descriptions dealt solely with the things in which the boys were interested, the order might indicate something about the interest of the boy who took the test. Burtt measured interest in terms of the ability of an individual to cross out irrelevant words in material which represented some special interest. Thus, he found that if he took material dealing with agricultural engineering and scattered through it a large number of irrelevant words, those men who were interested in agricultural engineering and who later made a success of it were not so successful in crossing out those irrelevant words. They became wrapped up in the subject and failed to notice the distractions by the way. He obtained a correlation of 0.30 with success in agricultural engineering. Time schedules showing attention actually given to poetry, detective stories, and biography are probably better indices of interest than formal statements would be.

t. Negativism in children is now being measured by a standard test situation. Small children are brought into the room, and rapport is established through a game. Children are then requested to pick up some blocks. Behavior is carefully noted.

After interest in a second game has been built up, it is interrupted with the demand that a pencil be picked up. Results of this study will probably be made available in a dissertation prepared by Miss Welty of Teachers College, Columbia University.

u. Persistence may be measured by the time during which an individual endeavors to solve any difficult problem. One investigator used an alley maze which could, by the readjustment of blocks, be made increasingly difficult and eventually impossible.

v. Recklessness

A test for taxicab drivers provided opportunity to take with a stylus a brief direct route through a maze, running the risk of a bad score through touching the walls, or an opportunity to detour through wider but longer paths. It seemed that recklessness might be revealed in the carelessness of the first choice.

w. Sociability

The children in a group may be asked to tell which children they would invite to a party, which are their best friends, etc. The extent to which an individual is chosen by his friends is surely the most valid possible measure of his popularity and social status. Other factors in sociability sometimes measured have been ability to recognize faces in pictures, knowledge of social terms, and amount of time spent at parties, dances, etc.

x. Social Perception

Ruckmick, Langfield, Laird, Remmers, and Gates have all had a part in the development of tests of the ability of persons to interpret photographs. The test of social perception, worked out by Dr. G. S. Gates, showed children pictures in which an actress was registering fear, anger, scorn, surprise, joy, and similar emotional states. It was found that ability to identify these grows with age and experience. Recognition of voice inflections from phonograph records in which accomplished actors recited the alphabet in varying moods, showed a similar course of development.

y. Studiousness

Studiousness has been measured in at least two ways. May asked students to keep a time schedule for a typical week near the beginning of the year and again near the middle of the year.

He found that, admitting the probability of exaggeration in report, the amount of time which each student reported as spent in study was a very good index of the kind of marks they would receive, an index fully as good as intelligence tests. Symonds suggests that a good way to measure studiousness is to use the ratio of the score (in S. D. units) on an unannounced quiz, and the score (in S. D. units) on a good intelligence test.

z. Suggestibility

A great variety of illusions, some of which are sensory, others involving more complicated conduct, have been studied to find out how far people can be influenced by suggestion.

One of the early experiments was to give an individual a magnetized ring which he was told would soon make his finger feel numb. A record was kept of the time interval until the numbness appeared. It seldom failed to develop. Other subjects have been asked to hold a wire and to report as the current was turned on, when the wire became too hot to hold comfortably. Many of them dropped it before any current had been turned on. The influence of size on judgments of weight has been frequently studied. Large objects are commonly counted as lighter than smaller objects of similar actual weight. One histrionically talented instructor suggested to a class that an odor which was produced by an experiment which he was conducting in the front was distasteful to many people, and sometimes resulted in their being overcome, and if any were very much bothered by it, they ought to leave the room. He found that numbers were badly overcome by the odor from the experiment from which no odor at all emanated. People can almost always, if they try hard enough, smell cloves or perfume, or fried onions.

The experiment of fidelity of report, illustrated above, might well be counted an experiment in suggestibility. When people are asked whether they saw a chair in the picture which was previously exposed, some persons are quite apt to remember one whether it appeared in the original image or not. In the Downey Will-temperament test there is a test of suggestibility. Subjects are given a true-false test with reference to a list of

words read some time before. As it happens, all of the statements are true, but they are told to correct them on the basis that half are wrong and half are right. Many persons will accept this statement as final and proceed to change their perfectly true answers. Voelker asked boys to make a choice, and then referring to it later said, "Which one did you choose?" "Oh, was it not the other? Just a minute, I made a record. Yes, it says the other." Some boys will agree with him, others will resent such suggestions. A test of suggestibility was included with the tests of decision types described above. Subjects were asked to indicate their preference among certain articles and activities, being previously told in each case, "Most people prefer this one." The statement about the one most people preferred was, of course, true only half of the time. Some subjects tended to respond to it, and to choose most frequently the one recommended by the apparent popular vote, others were influenced in a contrary fashion and tended always to avoid what had been chosen by most people.

6. Use of Standardized Tests

A final word should be said with reference to the use of any available test materials. Usually the directions in the accompanying manuals are very specific. They prescribe exactly the conditions under which the test should be given. It cannot be too strongly urged that these directions be followed just as they are stated. There is a particular temptation when dealing with groups in church schools, who are not accustomed to tests, for the leader to make an introductory speech. Not infrequently the originality of these introductory remarks injures the test results. Unless a test is given with precisely the directions with which it is supposed to be given, no words more or no words less, one cannot be sure that the results are comparable with those obtained when the test has been given as directed. Other temptations to irregular procedure may arise from the group itself. Persons not accustomed to rigorous schoolroom atmosphere will want to raise questions, perhaps make funny comments, and otherwise interfere with the test situation. A skilful leader will try to handle every such situation so that the total outcome would be

exactly what it would have been if the interruption had not occurred. If this is not practicable, notes should be taken as to just what the deviations in the procedure were, and the statement of result should be accompanied by this statement of error.

There is also a frequent temptation to alter slightly the directions or some of the statements in the test in order to fit better the needs of the particular group. If there is no desire to compare this group with any other, that procedure may be justified. The real value of a standardized test, however, lies in the fact that its reliability has been determined and norms have been established. Changes in directions or form of statements will almost certainly invalidate the norms. It is usually better to give tests in the form in which they have been prepared, making allowances, discounts, and explanations afterwards so that the test itself can be regarded as a fair measure on the old basis. Not every person who gives a test is in duty bound to improve it. In so far as he feels it possible to improve it, he may well gather those suggestions and send them to the author and publisher or may incorporate them in a new test. The business of test revision should be distinguished clearly, however, from the business of test administration.

a. Discussion of Test Results

Another problem is frequently raised by the Sunday-school classes and club groups which find the test questions so interesting that they want to discuss them. Many a leader having struggled for weeks trying to find problems of real concern to the group, gives them a test and is delighted to find that there is genuine warmth and interest in the questions raised. People want to know how certain questions are marked by other persons. They want to know the right answers for questions about which they previously had not thought. If the test is given only with the purpose of finding the present status of the group, it is quite justifiable to use it as a springboard into a series of discussions. The leader may write on the blackboard certain of the most interesting questions and tabulate the answers given by the group. Then the group as a whole may well consider why some people chose one answer while other people chose a different

answer. The principal caution in such a situation is that the leader should not spoil the educational value by a quick and dogmatic answer. One leader in the author's experience gave the Union Ethical Test and found that it raised a number of problems with the group. He saved 10 minutes at the end of the period and answered all of the questions. Other leaders in the same situation have found that such a group, in answering similar questions, could profitably spend as long as 6 months of study, investigation, and discussion.

If the tests are to be repeated at some later date, then discussion of the particular test elements must not be permitted. All tests are built on a sampling theory. Not every conceivable question about Bible information or economic prejudice is raised by any single test. Each one selects a certain sample from the total possibilities and judges what the general level of knowledge is by the range of achievement on the chosen samples. Now, if those particular samples are discussed, the level of score on the test is indeed raised, but it does not signify that the general level in the field which the test is supposed to measure has been raised correspondingly. Discussing the particular test questions is more or less like breathing on the thermometer. It raises the index, but it does very little to warm the atmosphere. In a situation in which the tests are to be repeated at some further time either in the same or in duplicate form, it is permissible to discuss the general topic. Thus, a class taking the Union Test of Religious Ideas at the beginning and end of the year, discovered that they knew little or nothing about the principal doctrines of religions other than Christianity. They took up this general topic, outlined the sort of information they wanted, and studied it without reference to any special elements which had been contained in the test. Another group, puzzled by certain questions regarding vocational choice, made that the theme of discussion for several weeks, inviting in special speakers to outline the preparation desirable in given vocations. In either case the test remained a fair index. The general levels of knowledge and understanding with reference to these fields had been raised through the discussion. The good effects were not

limited to the special questions raised by the test. Any progress which the tests showed at the end of the year could fairly well be interpreted as representing a real progress all along the line.

b. Cooperation in Standardization

In the present state of the measurement movement in religious education, while tests are in process of being prepared and standardized, authors very frequently send with the tests a request for cooperation. Sometimes the request is that the scores of pupils be returned, in order to help establish norms. Sometimes special report blanks are furnished. This practice seems to afford an admirable opportunity for scientific and Christian cooperation. No one who realizes the need for co-operative endeavor, if this relatively recent method of discovering and clarifying truth is to serve humanity as it might, will carelessly toss such requests aside. Every criticism and report which workers in the field can send back to the people who are particularly responsible for preparing and perfecting tests of all sorts will hasten the day when the number of measures in which we can have confidence will be adequate.

EXERCISES

1. Which of the tests now available might be used in studying your own problem?
2. Which of them would be most likely to be of service in connection with problem 34, page 258 in the Appendix?
3. Upon how many published tests are reliability figures available? How do these compare with the data on ratings given on pages 45–46.
4. Outline a life-situation test for tendency to dominate, to bully, to exercise mastery, as opposed to submission.
5. Outline a life-situation test for initiative.
6. Under what conditions is discussion of test results immediately after taking the test, desirable?

CHAPTER V

THE CONSTRUCTION OF TESTS

It is usually the case that, when an experimenter faces his own problem in a particular group, even so promising an array of tests as that set forth in the previous chapter fails to offer anything which is completely satisfactory. Some of the existing tests he finds that he can use, others he wishes to modify; there are still other elements upon which he wishes to equate groups or measure change, for which no test exists. Certainly it is true that for different age groups and different social environments different tests will be necessary even for the same trait.

The experience of test makers during the past decade has built up a number of guiding principles which may still be borne in mind by anyone who would construct new tests. Religious education is somewhat handicapped by the fact that many of these principles have been worked out in fields of public education in which subject matter seemed more important than it does to the person who conceives his task in terms of character formation. Most of the existing data upon reliability of different types of questions take it for granted that the purpose of the test is to measure information. These conclusions may or may not hold when the purpose of the test is to measure attitudes or mental and emotional processes.

The steps through which the experimenter will ordinarily proceed in the development of a new test are suggested below. Available experience upon each of these steps will be presented in the discussions throughout the chapter.

1. Analyze and define just what is to be measured.
2. Select the most suitable content.
3. Utilize the most appropriate forms for presenting this content.
4. Develop adequate directions.
5. Try out the test and revise it.
6. Develop a scoring scheme.
7. Standardize the test upon representative groups.

1. Definition of What Is to Be Measured

In test construction, as with almost everything else in experimentation, the motto should be, "Purpose First." Every step in the total process depends upon what the test is expected to do, and the groups in which it is expected to do it. Not all existing tests have been careful about defining purpose. They seem, in some cases, to be interesting collections of items that somebody would like to see tried out. Perhaps the satisfaction of the curiosity of the experimenter was the purpose. In any case, for careful scientific work they are about as useful as would be a futuristic painting of the persons studied. They may reveal many things but it is hard to know just what it all means.

Previous discussion has probably made it clear that "traits" are doubtful bases for test purposes.[1] They may be the best we have, but in the present state of knowledge it will usually be well to describe an honesty test, for example, much more specifically. Honesty in what sort of situations and under what conditions, is measured by this test? Is that the sort of thing that is significant for the purpose of the experiment?

2. Selection of Content

The next problem which confronts a test maker after he has chosen the field in which he wishes to construct his test and the type of attitude or process which he hopes to measure is to select the content material for his test. The usual process has been for an ingenious individual to sit down by himself and to evolve, out of his experience and hunches, a large number of interesting questions which might then be tried out, the poor ones eliminated, and the better ones kept and improved.

In many cases this process is capable of improvement if a preliminary step is taken. In some free and natural situation get an expression of the viewpoints with respect to the behavior, attitude, or philosophy which is to be tested. This may be done through group discussion in which the general topic is raised and careful note is taken of the positions which appeal to different

[1] See Watson, "Virtues *Versus* Virtue," *School and Society*, XXVI, p. 286, Sept. 3, 1927.

individuals. It may be done through oral interviews. A student who is formulating a test of worship attitudes carried on, first, a series of interviews with adolescents in which he raised questions about worship, and noted down all of their replies. It may be done through essay type questions. Allport, wishing to formulate a test of attitudes toward current political issues, secured themes from a large number of freshmen written on the questions upon which he wished to make tests. From these themes he was able to select phrases which were frequently employed and which embodied points of view which represented real possibilities for college students. Out of such discussions oral interviews, and essay type questions, it is possible to formulate controlled-answer questions which cover the ground far more adequately than do those which are concocted by any individual. A further distinct advantage is that the alternatives, when stated in the words of people who believe them, are likely to be much more appealing than when stated in the words of a test maker who does not sympathetically share the point of view he is trying to record. It is hardly necessary to state that the group from which the free responses are secured should be as representative as possible of the group with which the test is later to prove useful.

3. Form for Test Elements

The third problem is the matter of putting the questions in the most advisable form. Tests have usually been classified according to the form in which the questions are stated. The old-fashioned essay type of examination is familiar to everyone. A question is propounded, and pupils are given opportunity to select the ideas which they will use in answering it, to organize the material, and set it forth as best they can.

a. New Type and Essay Type Examination

Monroe, Starch, Gates, Wood, Ruch, and other investigators have studied the relative merits of the old essay examinations, and the new type recognition tests. One of the first differences noted was the difference in the objectivity of the score.[1] The

[1] These illustrations are taken from Starch, "Educational Psychology," The Macmillan Company, New York, 1919. Revised 1927.

same English paper given to a number of competent teachers in English was rated by some as low as 50 per cent, by others as high as 98 per cent. The marks on the same history paper ranged from 42 to 90 per cent, on a Latin paper from 45 to 100 per cent, and even on a geometry paper, supposed to represent an exact subject of study, the marks ranged from 93 per cent at the top of the scale, down to 28 per cent. The superiority of the new type examinations has been found to lie not only in the objectivity and reliability of the score, but also in such facts as these: they take less time both from the pupils who take the tests, and from the teachers who have to score them; they yield a finer grading of differences between individuals; they agree better with themselves, and are, therefore, more useful indices of a student's real ability; they are far more comprehensive because so little time is taken for each element; it is easier to make them diagnostic so that the spheres of difficulty or an individual's strong points and weak points can quickly be noted; they make it possible for comparisons to be made between classes and between different sections of the country; and they seem relatively free from suspicions of partiality in that each pupil can mark his own paper and recognize his own mistakes.

New type questions are usually of the controlled-answer type. They ask a question in such a way that a person can make only one of several possible answers. They become objective because any two persons given a score sheet would, when grading the same paper, come out with the same result. Among the common types of the newer examinations are true-false questions, two-answer questions, completion questions, word-phrase answer questions, and those which employ matching or pairing.

b. True-false Questions

The true-false question gives a statement, and the subject is asked to indicate by encircling the word "true" or the word "false," by putting a plus or minus, a "T" or an "F," or in some similar fashion, whether he believes the statement to be true or false. Provision should be made in the directions for a situation in which the pupil may believe the statement to be partly true and partly false. Usually the direction asks that such a state-

ment be marked false. In making true-false questions experience suggests need for caution at several points. One is the avoidance of extreme statements. "Dementia praecox always, in every case, arises directly from somatic sources" may not mean much to the reader, but any clever test taker would mark it false. Not knowledge of the subject but of the improbability of such a statement being completely true is sufficient for answering it thus. Another common difficulty is caused by the double negative which may arise in a false statement containing words like "no," "not," "none," and "never." This might be remembered best if dubbed the "Yes, we have no bananas" error. Still another common fault is that of combining with "since," "therefore," or "inasmuch" or even with "and" or "but" two clauses one of which is true while the other is false. Usually such a statement should be separated into two elements, each tested separately. Occasionally tests show poor form because the true-false statements are not clear and independent. In the endeavor to save time and space, elements are sometimes abbreviated, left as phrases instead of complete statements, or made dependent on some previous statement. This is poor technique. Each statement should be direct and complete. It is sometimes said that one difficulty with true-false questions is that they test only information. They may be used to test information, or they may be used to test high degree of insight into relationships, causes, and ability to make discriminations. Following are examples of some true-false questions used in one of the Y M C A tests.

ILLUSTRATION 1

Below are some statements, some of which are true, while others are false. Read each, and if you think it is wholly and unqualifiedly true, just as it stands, then draw a line under the word *True*. If you feel that it is not quite true, or is all false, then draw a line under the word *False*.

True *False* Parents who have plenty of money should give a boy as much spending money as he wants.

True *False* If a boy can get out of a scrape without telling the folks anything about it, he should do so.

True *False* If a boy has been forbidden to go to a show, and sees no good reason for it, it is all right for him to sneak out and go.

True *False* Parents have had more experience and know better what is good for boys, so boys should always do just as they are told.

True *False* A boy should be expected to do some things to help around the house, without being paid for it.

True *False* If a boy has only a very little money for an allowance, he shouldn't be expected to give any of it to church or charity.

True *False* Boys should tell their parents everything.

True *False* Usually a decision made after a boy and his parents have talked things over together is better than it would have been had either the boy or his parents made it alone.

True *False* A boy should spend all his spare time out of doors, getting exercise, playing games, and developing his muscles.

True *False* A boy should not join the church unless he believes everything the Bible and the creeds teach.

True *False* When boys are in school all week, Sunday should be a day of recreation and activity, rather than one of sitting still.

True *False* Jesus did not plan for churches as we now know them.

c. Two-answer Questions

Other two-answer questions may employ the classification of a statement as relevant or irrelevant, right or wrong, strong or weak, etc. The type of question is not very different from that found in true-false questions. The following are several samples. It will be noted that in some of them the direction is to encircle or check or cross out certain elements in a list. This is clearly a two-answer question because a person may either check the item or leave it out. It is equivalent to answering "yes" or "no," or "true" or "false." Some of the examples below show a modification of a straight two-answer type to include an intermediate answer or question mark.

ILLUSTRATIONS 2, 3, 4, 5, AND 6

2. Below are some of the arguments which have been advanced in discessions as to the existence of God. Read each and if you think it is a strong argument underline *Strong*. If you think it is not a strong argument underline *Weak*.

Strong *Weak* 1. Many qualities found in man could come only from God.

Strong *Weak* 2. Jesus assumed in all his teachings that there was a God.

Strong *Weak* 3. Men have always believed in God.

Strong *Weak* 4. Materialism is for most people an inadequate philosophy of life and socially an inadequate way of living.

Strong *Weak* 5. Man's whole nature receives some of its highest satisfaction through belief in God.

Strong *Weak* 6. Everything in the world works together for good.

Strong *Weak* 7. The entire universe seems best explained in terms of a God.

3. Below are the names of some men who have become famous. Judged by the standards of your course in "Studies in Leadership," which are truly great leaders? Draw a circle around the name of each man who, you think, was a truly great leader.

Julius Caesar	Asoka
Lincoln	Gautama Buddha
Roosevelt	Napoleon
Alexander	J. Pierpont Morgan
Grenfell	John R. Mott

4. Below are a number of questions, with possible responses after each. Consider each one briefly, and then, if you think the answer should be "Yes" draw a circle around the letter Y. If the answer should be "No," draw a circle around the letter N. If you are doubtful, encircle the question mark. Do not omit any.

Is the Bible best thought of as:

a. the correct and specific word of God? Y N ?

b. a library of many kinds of writings in many ages, by many kinds of persons? Y N ?

c. a record of what certain writers, mainly Hebrews, thought about life, long years ago? Y N ?

d. a series of principles and commandments, all of which we should follow today? Y N ?

e. an outworn collection of primitive science, legends, and customs with little meaning for modern life? Y N ?

f. presentations of the thoughts and ideals of some of the world's greatest religious leaders? Y N ?

g. studies of ideas and experiences which are important in modern life, as well as in ancient times? Y N ?

h. interesting reading? Y N ?

i. dry reading, but important? Y N ?

j. dry reading, and not important enough to bother much with today? Y N ?
k. some of the world's best literature? Y N ?
l. valuable primarily because it tells the life and teachings of Jesus? Y N ?

5. Below are listed four incidents, with two courses of action following each incident. Read each, and then answer the questions which follow it, by writing in either X or Y, which are supposed to stand for the names of boys.

Some children are playing out on the pier at a summer resort. A little girl falls off into deep water. X and Y are there, but neither one can swim.

X promptly runs and jumps in after her.
Y shouts for the lifeguard, who is not far away.
Which one was braver, in this case? _____
Which one showed a better-trained mind? _____
Which one did the better thing? _____

6. Below are some things which were true of Jonathan or of Solomon. Put (J) in front of all those which you think fairly describe Jonathan. Mark (S) in front of all those which fairly describe Solomon. If any statement is true of both Jonathan and Solomon put both (J) and (S) in front of it.

____ willing to give away something he wanted himself.
____ willing to risk his life for his friends.
____ ambitious for wealth and fame.
____ wise.
____ easily influenced by others.
____ successful in forcing others to obey him.
____ brave in battle.
____ obedient to God through his whole life.
____ sharing his wealth with God.
____ harsh with wrongdoers.
____ foolhardy.
____ loyal to his friends.
____ conceited.
____ self-indulgent and sensual.

d. Completion Questions

Completion questions are usually built in the form of paragraphs or long statements from which certain significant words have been omitted. They are, in many respects, the most difficult type of questions to construct. It is hard to make sure that there will be one and only one correct answer to be inserted

in the blank. There is no safe way to make completion questions without trying them out on sample groups. It sometimes happens that an ambiguous completion question will escape detection by a dozen people who examine the questions. Another difficulty in making out the completion question is to leave out words which are really significant. Certain words may be very easy to supply from the context, and hence afford no test whatever of the real point at issue. When completion questions are well made, they appear to be the most reliable tests of information which have yet been developed. It is very difficult to expand their use beyond information, to take account of attitudes and processes.

ILLUSTRATION 7

Every nation has its heroes about whom many stories and legends develop. Below are told in outline some of the stories about early Hebrew heroes. In each case some word or words are omitted. Read each story and write in the word or words which will best complete it.

Once the king of a certain country ordered all the boy babies to be _____. The baby Moses was hidden in_____ _____and rescued and taken home by_____. When he grew up God wanted him to lead his countrymen out of this country of _____. At first Moses _____ _____ but later he _____.

After the death of Moses _____ became leader. This leader led the people across the _____ River and conquered the new land.

Another hero about whom many tales are told was Gideon. One of the first things we know about him is that he_____ the altars in his home town. As commander of the army, he chose only _____ men and with these defeated the Midianites. When the people wanted him to be _____, he decided _____.

e. Word-phrase Answer Questions

The word-phrase questions are very much like completion question except that they usually call for the writing in of only one key word or answer. Most arithmetic problems are of this type if answer only is taken into account. The answer may be very brief, but it is the result of a long train of thinking which has

preceded it. The following example shows a word-phrase question in which names are used for answers.

ILLUSTRATION 8

Below are some stories you have studied, with the names omitted. Read through each story here and fill in the name of the man to whom it belongs.

_____, who was a publican, made a feast at his house for his friends and Jesus that they might know the man whom he was following.

_____, a wealthy young man, sold his land and gave the proceeds to the church. Later he joined Paul in some of his journeys and we find his name associated with deeds of generosity and of aid to those in distress.

_____, a physician, accompanied Paul on many of his journeys and wrote the story of them.

After the persecution following the death of Stephen, _____ _____ went down to Samaria to teach and, while there, met and converted an Ethiopian who was trying to read and understand the Old Testament prophecies.

f. Multiple-choice Questions

The most frequently used questions in tests on religious ideas and ethical judgment are multiple-choice questions. Usually a situation is stated and several possible ways out are suggested. Sometimes the individual is asked to pick the best one of those answers. Multiple-choice questions tend to be easier than other types of questions, largely because the variety of possibilities encourages the examiner to put in suggestions which are obviously not the best. They are not live options for the person taking the examination. Many multiple-choice questions which appear to have five choices really have only two or three. The rest may just as well not be there. They are a waste of printing space. It is a safe rule not to include in a multiple-choice test an option which was not checked by anyone among the first hundred sample individuals tested. There is another error in the construction of multiple-choice tests which runs in the opposite direction. Sometimes the answers are too close together. It becomes almost impossible to judge among them. The shades of difference are so slight that almost any answer can be justified. This type of question is also open to criticism. Sometimes

makers of multiple-choice questions tend to put the correct answer first or last or in some other constant position. The arrangement should, of course, be purely random. The multiple-choice question is apt to consume a great deal of space in printing, especially when a number of possibilities are listed and only one is to be checked. It is possible to alleviate this somewhat by stringing out across the page the answers to be used, rather than listing them, or using the same material more than once, that is, having one sort of mark for best and another sort of mark for poorest. In a question like Illustration 9, scoring is made easier if the possible answers are numbered, and a blank provided in the margin, in which the number of the correct answer is to be written.

ILLUSTRATIONS 9 AND 10

9. In the following sentences the last word or phrase is omitted. Several possible words or phrases are given, one of which is the correct one. Read them carefully and underline the one that makes the best answer:

a. Before his ministry began Jesus was a plumber barber carpenter salesman teacher.

b. Jesus drove the money changers from the temple because they were Jews it showed people he was brave they were desecrating the temple they were getting rich.

c. Jesus performed miracles of healing the sick to win popularity to show His power to relieve suffering to show He was divine.

d. Jesus says your neighbor is one who lives near you needs your help has helped you before.

e. The man who had been beaten and robbed was helped by a foreigner a church official a teacher a policeman.

10. Most people agree that swearing is an undesirable habit. Below are reasons sometimes given in support of this view. Read them all carefully. Then make a check mark in front of that one which seems to you the best reason of all. Check only one.

____ the Bible says it is wrong.
____ it is a bad example for smaller boys.
____ it is an expression of strong emotion, and all violent emotion is undesirable.
____ it is unclean.
____ it develops the habit of loose thinking.
____ it shows a lack of self-control.

Now read these reasons all again. Cross out that one which you think the poorest and weakest of all. Cross out only one.

g. Degrees Type of Question

Another type of question which looks quite different from the multiple-choice really embodies the same principle. That one allows an individual to classify a statement as being in one of three or four or five classes, that is, it may be regarded as excellent, good, fair, poor. It may be attributed to many, few, or no persons. It may be rated on a scale from one to ten in importance, or in some similar fashion the judgment of the subject with reference to the statement may be expressed. It is really a multiple choice in that he chooses into which category or degree he will place the question. This type of multiple-choice question has a distinct advantage over the previous type in that every statement is reacted to. In the multiple-choice question in which only the best is checked, there are two or three or four or more upon which the subject makes no comment. It is assumed that he thinks them not so good as the best one, but in this multiple-choice form which employs the degree of truth or desirability in each suggested answer a reaction is made to every statement.

ILLUSTRATIONS 11, 12, 13, 14, AND 15

11. The Hi–Y group at one of its meetings discussed prayer. A number of the statements below were made. Consider each one, and if you wholly agree with it, feeling it to be an excellent statement about prayer, draw a line under the word *Excellent*, in the margin. If you feel that it is only partly true, or is only a fair expression of what you really think, then draw a line under the word *Fair*. If you don't believe it, or feel it to be a poor, or unimportant statement, then draw a line under the word *Poor*.

Excellent *Fair* *Poor* 1. Prayer is a court of last resort, to be used only after all other methods have failed.

Excellent *Fair* *Poor* 2. We all need to pray regularly at some fixed time and place.

Excellent *Fair* *Poor* 3. Prayer gives an opportunity to think things through with God.

Excellent *Fair* *Poor* 4. Few of us pray as often, as expectantly, or as intelligently as we ought.

Excellent *Fair* *Poor* 5. It is better to do the thing than just to pray about getting it done.

Excellent *Fair* *Poor* 6. We should work as though everything depended on us, and pray as though everything depended on God.

Excellent *Fair* *Poor* 7. An inner-circle group, just by praying, often completely changes the spirit of a school.

Excellent *Fair* *Poor* 8. Prayer is a waste of time.

Excellent *Fair* *Poor* 9. Prayer makes you dependent on another person instead of being independent and self-reliant.

Excellent *Fair* *Poor* 10. We should all pray because prayer has been reported to accomplish miracles.

12. Below are a number of statements with the first word omitted from each one. You are given a choice of five words to use as the first word of the statements: "all, most, many, few, no." Indicate which word you think makes the best statement by drawing a line around one of the words before each statement. Draw a line around just *one* of the words.

All *Most* *Many* *Few* *No* fellows are capable of choosing their vocation wisely by themselves without help from others.

All *Most* *Many* *Few* *No* young men who have skilful advice and choose their vocation intelligently get into the work for which they are best fitted.

All *Most* *Many* *Few* *No* boys who decide early in life (12 to 14 years) on their life work change to some other vocation before they actually start to work.

All *Most* *Many* *Few* *No* men with college training make more money than those without such training.

All *Most* *Many* *Few* *No* young fellows should go to college to prepare themselves for their life work.

All *Most* *Many* *Few* *No* boys who start to work between 14 and 16 years of age get into "blind-alley jobs."

All *Most* *Many* *Few* *No* young fellows can find an occupation which is desirable from every point of view and has no drawbacks.

All *Most* *Many* *Few* *No* occupations are suited to all fellows.

All *Most* *Many* *Few* *No* men determine their vocation by chance rather than forethought — drift into it.

All *Most* *Many* *Few* *No* men who drift into their vocations get into one for which they are really suited.

13. Below are several statements. Read each, and if you feel it to be wholly and unqualifiedly true, place a check in the column under the words *Wholly True.* If you feel it to be frequently true or true in many circumstances, place the check in the column under the words *True in Large Degree.* If you feel that it is quite uncertain, place the check in the column under *Doubtful.* If you feel that the statement is largely or entirely false, place the check in the column under *False.*

Wholly True	*True in Large Degree*	*Doubtful*	*False*	
____	____	____	____	1. God's hope for the world is the same as that of intelligent Christian men.
____	____	____	____	2. Prayer offers the only way of finding out God's hope for the world.
____	____	____	____	3. Knowledge of God's hope for the world will spring from an interest in world conditions today.
____	____	____	____	4. It is a real question whether civilization will go to smash, or will become Christian.
____	____	____	____	5. The hope of the world lies in trusting our leaders.
____	____	____	____	6. The world today is moving rapidly toward the fulfilment of God's hope for it.

14. Below are some things which might have something to do with being healthy. Read each one and then if you think that it is a thing which is absolutely essential for being healthy, put a check (√) under "absolutely essential." If you think that it is "very desirable," put a check in that column; if you think that it is "unimportant," put a check in that column; or if you think that it is a thing that would harm your health, put a check under "harmful."

Absolutely Essential	*Very Desirable*	*Unimportant*	*Harmful*	
____	____	____	____	Getting at least nine hours' sleep every night.
____	____	____	____	Eating simple food.
____	____	____	____	Going to shows.
____	____	____	____	Hiking.
____	____	____	____	Eating some candy every day.
____	____	____	____	Sleeping with windows open.
____	____	____	____	Smoking.
____	____	____	____	Playing pool.
____	____	____	____	Being out of doors.
____	____	____	____	Dancing.
____	____	____	____	Having regular habits of elimination.
____	____	____	____	Petting.
____	____	____	____	Drinking three pints of water a day.
____	____	____	____	Playing basketball for several hours four nights a week.

15. People vary widely in their opinions of the activities listed below. Indicate how you feel about each by writing in front of it a figure from the following scale.

+ 3. Sure to be desirable, worth while, enjoyable.
+ 2. Usually all right, interesting, worth while.
+ 1. Sometimes all right.
0. Don't know, makes no difference, can't say.
− 1. Sometimes undesirable.
− 2. Usually harmful, undesirable, uninteresting.
− 3. Sure to be undesirable, harmful, bad, etc.

____ 1. Having girl friends.
____ 2. Smoking a pipe.
____ 3. Saying "damn" or "hell."
____ 4. Buying stocks, hoping to sell when they rise.
____ 5. Reading poems.
____ 6. Reading lives of great men.
____ 7. Buying five sundaes a week.
____ 8. Daydreaming about success in life, high honors, etc.
____ 9. Going to Sunday school.
____ 10. Working on a committee in religious work of some kind.

In some similar fashion the multiple-choice may be combined with the true-false question. An incident is stated, the answers among which choices are to be made listed, and then instead of being asked to choose only the best, the subject is asked to rate each one of the proposals as right or wrong, strong or weak, true or false. An example follows:

ILLUSTRATION 16

Several life situations are told below. Read each one carefully and if you think it would be fair to make such a judgment, draw a line around *Yes*. If it would not be a fair judgment to make, then draw a line around *No*.

(*a*) Two boys must work after school hours to help pay their way. They are good basketball players but are not playing basketball. Some students talk about them and attempt to make them unpopular by saying they are not loyal to the school.

Yes *No* 1. The players should give up their work.
Yes *No* 2. The boys who questioned their loyalty had little appreciation of the difficulties of poverty.
Yes *No* 3. The loyalty of the boys who talked was a false loyalty.
Yes *No* 4. The boys who talked simply used unwise methods.

(*b*) A group of underpaid workers having been refused a "raise" remained at work but worked as slowly as possible and turned out imperfect articles.

Yes *No* 1. The responsibility for this should fall upon the employers.
Yes *No* 2. The workmen were unfair to the people who bought the goods.
Yes *No* 3. The workers are justified in using their power in their jobs as long as the employers use the money power they have.
Yes *No* 4. Both employers and workers were wrong.
Yes *No* 5. The public was partly at fault in this case, because they were willing to buy goods made by underpaid workers.

(*c*) A minister in a mining town urged his congregation to supply food for the needy families of some strikers.

Yes *No* 1. He was taking an unjustifiable part in a fight between the strikers and the mine owners.
Yes *No* 2. He was doing a humanitarian thing which did not mean that he was necessarily on the side of the strikers.
Yes *No* 3. He was outside of his field of work in urging this.
Yes *No* 4. He was practicing the true gospel of Jesus.

(*d*) A wealthy banker's son and a janitor's son were chums. As they grew up the wealthy parents tried to discourage the friendship.

Yes *No* 1. This was wise because they would have to go with different social sets when they grew older.
Yes *No* 2. The banker's son should have refused his parents' wishes.
Yes *No* 3. The janitor's son should have broken the friendship.
Yes *No* 4. There should have been no distinction between bankers and janitors because of their positions.

(*e*) In a men's Bible class one man said that 30 per cent of all workers were unwilling to work any more than they had to and for this reason deserved to be poor.

Yes *No* 1. He was correct in his conclusion.
Yes *No* 2. It was a brutal, unsympathetic remark.
Yes *No* 3. He failed to take account of the conditions that made these people unwilling to work.
Yes *No* 4 He should have blamed the big money interests.
Yes *No* 5. He could be of this opinion and still be Christian if he was willing to help any of the 30 per cent who reformed.

(*f*) A student in an eastern university has three thousand dollars a year to spend and one thousand to give to charity.

Yes *No* 1. He is likely to think he is better than other people.
Yes *No* 2. He is likely not to develop a strong character so that his value to society will be less than that of a fellow who has less money.

Yes *No* 3. He cannot be sympathetic with poor people.
Yes *No* 4. He should give away most of his money and work for part of his expenses.
Yes *No* 5. He has an unusual chance to render significant service.

h. Ranking Questions

Ranking questions are theoretically multiple-choice, but they practically are somewhat different in the type of action they demand. The subject is asked to arrange in order of goodness or aptness or truth or preference certain statements or descriptions. There may be only two of them as in the first statement below, or there may be as many as ten. In practice it appears that children in grade schools cannot effectively bear in mind more than five proposals that they are to rank. Ranking, of course, involves comparing each one with every other one. Ten is as many as an adult can handle comfortably. If there are more to be ranked it is better to break them into several groups of ten each, carrying over one element which is common in every group if it is desired to make them comparable. Again in ranking questions it is necessary to watch carefully the two dangers of multiple-choice questions: The first that suggestions may be so unusual that no one would ever seriously consider putting them anywhere except at top or bottom; the second that some may be so close together that it is impossible to decide clearly and fairly which should come first. If this latter danger cannot be wholly eliminated, it is possible to provide for it in scoring by permitting a certain range of answers. Thus, it may be that when it comes to scoring on the ranking questions with ten elements, either of two would be permitted to be ranked first or second. Any of the next five would be credited as correct if ranked third, fourth, fifth, sixth, or seventh, and any of the remaining three would be counted correct if ranked eighth, ninth, or tenth. The following are examples of ranking questions with two, four, and ten elements.

ILLUSTRATIONS 17, 18, AND 19

17. Below are listed several groups of alternatives. Read each pair and place a check in front of the one of the two which seems to you to be more Christlike.

____ There will be greater happiness if the more capable are allowed to accumulate great wealth and endow parks, libraries, schools, etc.

____ The greater happiness will come if society restricts the accumulation of the few and gives everyone a chance to attain and contribute to the welfare and happiness of the common good.

____ Boys and girls should have an equal amount of freedom in sex matters as regards such things as petting and the like.

____ It is harder for boys to control themselves, and they should not be expected to live up to the standards that are set for girls.

____ Owners of large business concerns should see that all their employees have a fair wage, but should keep the management in their own hands.

____ Capitalists and laborers should own and run a business in common.

____ Christians should try to convert all nations to their belief.

____ "Heathen nations" should keep the good things in their faith and only take the things from us that will enable them to make it better.

____ A nation should refuse to fight but should forgive a nation which has treated it unfairly.

____ A nation should go to war to get even with another nation which has treated it unfairly.

18. Below are four instances and following each instance are four ways of acting. In each case read all four ways of acting and put a (1) in front of the one you feel to be the best, a (2) in front of the next best, a (3) in front of the next best, and finally a (4) in front of the one that you think would be the poorest way of acting.

A boy has taken his older brother's fountain pen and kept it.

____ He should be severely punished.
____ He should be given a chance to earn one for himself.
____ He should be lectured severely.
____ He should be given one of his own.

A poor man breaks into a store and steals some money.

____ He should be severely reprimanded by the judge.
____ He should be given a pension to help support his family.
____ He should be helped to find a job.
____ He should be sent to jail.

A member of the opposing team has slugged you.

____ You should fight him after the game.
____ You should pay no attention but play your best.
____ You should report him to the umpire.
____ You should slug him every chance you get.

A member of the office force has misrepresented you to the boss, saying that you were discharged from your last job because you were suspected of having taken some money.

____ You should pay no attention but do your best at your work.
____ You should spread reports around about the other fellow.
____ You should tell this fellow what you think of him.
____ You should go to the boss and tell him the truth.

19. In the instance below *rank* the suggested answers. Place (1) in front of the best, (2) in front of the next best, (3) next, then (4), (5), (6), (7), (8), and (9) in order, placing (10) finally in front of the worst or poorest suggested answer. Be sure every one is ranked.

Rank the following methods of keeping Sunday, from the standpoint of their importance to a present-day Christian boy.

____ *a.* not making any noise.
____ *b.* washing dishes for your mother.
____ *c.* going to church or Sunday school.
____ *d.* reading the comic section of the Sunday paper.
____ *e.* playing on a community baseball team.
____ *f.* praying a great deal.
____ *g.* eating a big dinner.
____ *h.* wearing your best clothes.
____ *i.* reading a good book.

i. Matching and Pairing Questions

Matching and pairing questions have proved very useful in connection with subject-matter quizzes. Usually one set of ideas is given, each of which is to be paired with some idea in a second set. Sometimes these are names and dates; sometimes they are phrases and books of the Bible; sometimes they are cause and effect. The illustration below is taken from the field of Christian biography.

ILLUSTRATION 20

In the first column below are the names of certain leaders; in the second column are listed ideas, deeds, or movements with which the leaders have been connected. After the name of each leader write the number corresponding to the idea, deed, or movement which is connected with his name.

Jonathan Edwards ____	1. The Reformation.
Savonarola ____	2. Rise of Methodism.
Benedict ____	3. Stirring the city of Florence.
Constantine ____	4. Writing his Confessions.
Thomas Aquinas ____	5. Formulating an intellectual, logical defense of the medieval church.

Wesley ____	6. Giving official recognition to Christianity.
Calvin ____	7. Scotch Presbyterianism.
Channing ____	8. Unitarianism.
Augustine ____	9. Freedom of the will.
Luther ____	10. Founding a monastery.

4. Other Significant Classifications of Test Elements

When the purpose of a test is to find out how adequately a pupil is informed in any given field, the matter of form of question is one of the chief variables for study. In our present situation in religious education it is doubtful whether form is the principal variable to be considered. It probably matters far less whether the questions are in true-false or multiple-choice form than it does whether the question asks for an opinion or calls for an involved comparison. We may look over the questions in existing tests from many other points of view than simply the point of view of form of question.[1] For example, all tests may be

[1] Hartshorne and May, in their article "Objective Methods of Measuring Character," *Ped. Sem.* and *Jour. Gen. Psy.*, March, 1925), classify the existing tests under the terms:

1. Order of merit.
2. Scale of values.
3. Multiple choice.
4. True false.
5. Cross out.
6. Distraction.
7. Information.
8. Completion.
9. Recognition.
10. Preference
11. Association.
12. Physiological.

They suggest another classification in terms of the situation in which the subject was placed and the response expected of him. The situation then might be:

A. Natural uncontrolled.
B. Natural controlled.
C. Experimental and controlled.

The response might, in similar fashion, be:

A. Natural undirected.
B. Natural directed.
C. Experimental directed.

Any combination of these is possible.

classified according to the degree of indirection with which they approach the problem of behavior. Thus, we would have:

1. Tests which ask pupils what they would do in given situations.
2. Tests which ask pupils what they would think it best to do in certain situations.
3. Tests which ask what other people would be most likely to do in certain situations.
4. Tests which ask for indirect clues, hoping that these will give some guidance as to what a pupil would really do without his being aware that he is answering that kind of question. Thus, the question that asks for his preference among chums may really be testing which interest he would be most likely to take wholeheartedly, or tests which discover his information about baseball as compared with popular music giving a fair index as to how he would spend some spare time.

It seems to be true that the first of these types, the question which goes directly at the matter and asks what a person would do, is of comparatively little use in the realm of ethical discriminations. A good answer may equally well be made by a person with high ethical development and by a person who is a hypocritical liar. Any test which groups together in indistinguishable fashion these two traits will not long be satisfactory. Most of the questions in the illustrations above belong to the second type. They strive to find out what the subject believes it best to do. Such tests eliminate rather clearly people who do not know what it is best to do in social, ethical, and religious situations. They do not necessarily separate those who would really do the best thing from those who would know it but not do it. The third and fourth types, because they tend not to put an individual on his guard, are somewhat more likely to give real information as to what a person would do.

Again tests may classified with reference to the intellectual processes upon which they call. We can distinguish among existing questions such types as:

1. Tests which ask a pupil to remember certain items of information.
2. Tests which demand that the pupil express a preference.
3. Tests which require an insight into consequences, the usual cause and effect relationships which prevail in human life.
4. Tests which require understanding of one's fellowmen to the extent of knowing what their preferences usually are and what their behavior in certain situations is most likely to be.
5. Tests which place a premium upon critical mindedness, a tendency toward wariness when confronted with catch phrases and platitudes of either the church or the poolroom.

It is possible to classify tests again with reference to their content and the place where they seek to test an individual's attitudes and responses; for example, we may have:

1. Tests of attitudes and behavior which lie largely within the life of the individual, tending to affect no others beside himself.
2. Tests employing situations within the family.
3. Tests employing attitudes toward the church and behavior within the church organization.
4. Tests based on school life.
5. Tests dealing with the work group, choice of work, industrial justice, and the development of Christian economic relationships.
6. Tests of attitude within the play group, evaluation of recreation.
7. Tests within the national and international relationships.

At any of these levels the test may be predominantly immediate in its type, in which case we may think of it as primarily ethical or moral or it may endeavor to bring into the problem the total attitude of an individual toward life and the universe, in which case it becomes more nearly a religious test. These questions might be thought of from still another point of view as being tests which show conformity with the existing standards of society, or as tests which demand a creative reconstruction of present accepted standards. There is need for special concern

over the tendency of most "ethical" tests to accept uncritically the customs of modern society.

The principal point of emphasis in these many but surely not exhaustive attempts to classify existing test questions is that the study of form is likely to be overdone. It is probably more important to know the differences in reliability and validity of questions on the basis of the other classifications suggested here than it is to know the differences on the basis of true-false, completion, etc. In the present state of our knowledge, the particular form of construction of the test questions will matter very much less to the person taking the test and to the interpretation which can be placed upon it afterward, than will the question of the point at which the test falls when its content is considered and the processes upon which it calls are analyzed. It is urged that students of measurement in religious education begin forthwith to study test elements as carefully as possible from the viewpoint of new and more fruitful classifications.

5. Development of Directions

Suggestions have thus far been made with reference to the choice of material and the choice of form in which the material is to be arranged. It is perhaps incidental to the matter of form, but so important as to merit separate treatment, to consider the formulation of directions. Directions should provide a motive for taking the test. In so far as possible they should tell the truth, or at least as much of it as will give the subject the most desirable attitude during and after the examination. Directions should be brief, avoiding repetitions; directions should be complete; directions should use simple words; directions should be broken into action units. If the first thing to be done is to read a certain paragraph, then that should be stated. After the paragraph is read the next direction should be given. If it seems particularly desirable to bunch the directions together, the order of instruction should be the same as the order of execution. Ordinarily, it is better to give only one set of directions at a time. Thus, in example 10 above, where the best and worst were to be chosen, pupils were asked to take only the best when they considered the problem the first time, and

were at the end of this work instructed to go back over it again and pick out the worst. Ordinarily directions should contain an example or practice exercise. This is particularly necessary in the case of pupils who have not taken tests of the kind previously. Moreover, tests differ widely in the precise type of response which they require. Some true-false tests use plus and minus, some tests use yes and no, some ask that you circle the perfect word, others that you cross out the words that you do not wish to leave standing. Pupils become very easily confused unless they have a chance to do a practice exercise first, to ask any question about it which they find it necessary to ask, and so become ready to consider the real test without the handicap of worrying about the form of response.

Different types of question require certain special directions. In the true-false question it is necessary to make some provision for the statements which seem neither true nor false. In the multiple-choice question it is necessary to make especially emphatic whether only one answer is to be chosen or whether several may be included. Completion directions are usually simple, but ranking questions are very apt to be misunderstood. Considerable experience has shown that with persons below high-school age and sometimes with older persons, it is necessary at the beginning of a ranking question to describe very specifically just what is to be done. On the Y M C A examinations one year, the directions went something as follows: "Rank the acts given below, placing 1 in front of the best, 2 in front of the next best, 3 next, etc., finally placing 10 in front of the poorest." There were at least a score of papers on which only four alternatives were marked: 1, 2, 3, and 10. The other numbers not being present in the directions were not understood by the mere words "rank" and "etc."

If the scoring system is to be in any way unusual, directions should take account of this fact. Directions should make it perfectly clear whether an individual who does not know should guess or should leave the statement blank. If there is to be a penalty for guessing, it should be stated.

After the test material has been selected, formulated, and

directions made as clear as possible, the test is ready for its preliminary trial. It should be given first to a comparatively small group of representative persons who should be encouraged to note down by the way anything which they find confusing or which they believe does not contribute to the purpose of the test. Tests may frequently be given to seminar groups of graduate students, after which the tests may be discussed. This has proved to be a very fruitful method of testing elements and form. It is very easy for a test maker to resent criticisms at this stage, to feel that his work so far has required so much time and energy that he is in duty bound to defend this test in its preliminary trial. Sober thought makes it clear that all criticisms should be welcomed, and elements changed in the light of new knowledge if the test is really to serve a wide group. The maker of the test may not feel about his work of art as the maker of a poem or drama might feel. If no one likes the poem but the poet, that is sufficient. The same can hardly be said of useful religious tests.

6. Methods of Determining Scoring Scheme

The next major problem is the determination of the scoring scheme. Probably the poorest method is to use the judgment of the individual who made the test as to what answer he would like to have the people make. There is some improvement over this method when the test is submitted to a large number of experts in the field and their judgment is taken as representing the best answer. The very selection of the experts, of course, tends to determine the type of answer that will be given particular consideration. In matters like a Bible history test this is quite justifiable. Expert knowledge can be checked against reliable sources of information. Some answers are more truly right than are other answers. When it comes to matters of ethical choice and personality make-up, it is not so clear that any group of individuals can select the best answer. The difficulty becomes still more aggravated when the test is to be given to small children because the best answer for an adult is frequently not the best answer for a child.

There are two outstanding methods for determining the best scoring system. One is based upon the assumption that true

answers can be obtained, the other is based upon the assumption that specific answers are difficult to determine, but that it is possible to select people possessing more or less of the general trait which is being measured.

a. Scoring when Best Answers Are Known

When dealing with questions of Bible information and other items upon which valid answers can be obtained, there still remain several methods for awarding credit. The simplest is to give one point for each right or significant response. A second method is to correct for errors due to chance and guessing. Thus, if the tests are of the true-false type and there are 50 questions, it is clear that a person who was completely ignorant with reference to the material being tested, but who could understand the directions, would, on the average, have 25 right and 25 wrong. The real measure of his ability should not be 25 but zero. On this theoretical basis, the scoring system for two-answer questions should be "right minus wrong" rather than simply the total number of rights. The assumption here is that for every error which a person has made, he has made two guesses, one of which was wrong, the other of which was right. Studies of the results of scoring systems have not been entirely conclusive. Within the same month two articles appeared in technical journals. In one of them the writer said, in substance, "I have always previously recommended that a correction for guessing should be made. On the basis of present study of the reliability of tests, I am abandoning this practice because the uncorrected result is slightly more reliable than the corrected result." The other writer said, "I have always heretofore recommended that the total number of right answers be regarded as the score in a true-false test. I find upon study of actual test results that when a correction for guessing is made, the results do become more valid than they were previously. I am, therefore, adopting the policy of instructing subjects not to guess, and then penalizing for guesses." In the realms with which most of the religious and ethical tests deal, it is very difficult to separate a guess from an opinion or a preference. The recommendation of the writer is that the pupils be instructed to answer every question, and that the total

number of right answers be taken as the score. This yields precisely the same relative order of score for the pupils, if they have answered each item, as would be given if the correction process, which is more difficult to justify, were used.

A third method of expressing a score when the correct answers can be defined is to use not simply the number right, but the per cent right.

A fourth method for expressing the score of a pupil is the use of percentile units.[1] On some basis the actual number of right responses is established, then a distribution is made in which the highest scores are placed at the top, then the next to the highest, and so on down to the lowest scores which stand at the bottom. The percentile score at any point expresses the per cent of the remaining subjects who receive a lower score. Thus, the lowest person receives a percentile score of zero; the best one receives a percentile score of 100; the average pupil receives a percentile score of 50. The pupil with a percentile score of 80 knows that he has done better work than 80 per cent of the people who have taken the test, but that 20 per cent have surpassed him.

A percentile unit is frequently criticized because it seems that the difference between a score of 50 and a score of 55 is the same as the difference between a score of 92 and a score of 97, whereas in almost every ability which has been measured, the latter is many times as great as the former. Out near the extremes it is harder to surpass ten people than it is to surpass ten people taken near the middle of the scale. Near the middle, abilities seem to be bunched more closely together. A certain improvement might move one ahead of 1,000 average persons, whereas it would move one ahead of only one or two if the competition began at the high end of the scale. This can be corrected by the use of the standard deviation unit for expressing a score. McCall in his "How to Measure in Education," presents a scale based on the standard deviation which he calls in honor of Thorndike and Terman, the T scale. The units which he uses are one tenth of the standard deviation of the distribution. His zero point is taken five standard deviations below the average

[1] See p. 188–89.

of the group. This means that a T score of 50 represents average achievement for the group studied (McCall used 12-year-old pupils in public school) whereas a score of 100 would be a perfection almost never attained, and a score of zero the degree of failure not met once in 10,000 cases. If the trait being measured follows a normal distribution as most human abilities seem to, then one can safely say that the difference between a score of 50 and one of 55 in S. D. units is no more and no less than the difference between a score of 0.92 and one of 0.97. TableXXXIII on page 278 gives the percentage of the total number of cases, making scores which are higher than any indicated score in S. D. units. Thus, there are 50 per cent who have scores above an S. D. score of 50 and 1.07 per cent who stand above an S. D. score of 73, while 99.2 per cent or practically all of the group tested might be found above an S. D. score of 26.

A sixth method for expressing a score when the right answers can be definitely fixed is in terms of age or grade norms. This method has become familiar through the practice of intelligence testers. A child's mental ability is usually not stated in points or in percentile units or in S. D. units, but in terms of ability corresponding to the average child of eight or ten years. School achievement is frequently expressed as being equivalent to fourth-grade or seventh-grade level of ability. Present work on the correlation between age or grade on the one hand, and ethical and religious qualities on the other, seems to indicate that these methods will not be useful in connection with ethical tests. There are some very young boys and girls who do quite as well as do others who have had much more school training and much more experience in life. The use of the terms "moral age" and "religious age" seems to be more dangerous than promising.

b. Development of Scoring on Basis of General Criteria

The second general situation demanding a scoring device is one in which it is difficult to choose particular responses as right or best, but it is possible to choose persons having various degrees of the general trait it is desired to measure.

In this situation a method has been developed for determining the best answer to a test question in terms of what it really

indicates. The method may be illustrated most readily in the realm of vocational tests. Suppose it were possible to have a foreman and other capable persons in an industry select those workmen whom they would consider worthy of a "double A" rating from the standpoint of efficiency, those who may be considered fairly good, another average group, a group falling below average, and a fifth group of very, very poor individuals. Suppose it is desired to make a test which will, in advance, separate one group from another and predict correctly which individuals are most efficient and which ones belong in the middle of the scale or at the bottom. The test may be tried out on the total group and then studied to find out which elements were answered correctly by the highest group but by no others, which elements were answered by excellent and good groups but not by the average or poor, and so on down the scale. Certain errors, it will be found, will be made only by the poorest group. No answer, however good it might appear to the test maker in advance, which was answered correctly as often by the people in the lowest efficiency group and people in the highest efficiency group would be a good test element. It would not differentiate.

In the same fashion scoring systems can be built up for tests in other spheres. Suppose we can establish by means of other tests, interviews, ratings by competent persons, and in other ways, that group *A* is composed of people who are devout, thoroughly worshipful individuals; that group *B* is composed of average persons from the standpoint of their interest in worship, and that group *C* is composed of cynics, scoffers, persons whose chief delight is in showing themselves irreverent upon all possible occasions. We could then try out a test for worship participation and find out element by element which ones differentiate one group from the others. If we find some elements which are answered in one way by half the people in *A*, half the people in *B*, and half the people in *C*, it is clear that those elements are not of much use.

The first step in this process of determining the score to be given to test elements is to build a criterion. We must have some basis for knowing which people ought to have high scores on the

test and which people ought to receive low scores. In some cases this may be done by giving a large battery of tests. If we were to give all the existing ethical tests to a wide group we might have some basis for saying that certain people belong at the top of the scale and certain others at the bottom. Usually existing tests are not enough. The criterion must be built up by the use of ratings, judgments of intimate friends, case studies, reports of behavior, action, and conduct tests. The group used as a criterion should so far as possible be representative of all the types of persons who will eventually take the test, but it should be small enough so that each member of the criterion group can be studied individually. Sometimes it happens that there are no tests or ratings in which one has a great deal of confidence with reference to a particular character trait. It may be that the test itself is the best evidence available. Suppose that an advanced Bible test is being studied from this point of view. It may well be that no one is more competent to determine which individual should stand high and which one should stand low than the person who has given the test which is being created and has scored the answers from his own point of view as to which ones are wrong. In such case, the total score of the individual, using a scoring system built out of the best available judgments, is utilized as the criterion.

c. Technique of the Criterion Score Process[1]

Following are the steps in the process of determining the weight to be given each element:

1. Build the criterion.

2. List across the top of a sheet every possible response to every element in the test.

3. Tabulate the answers for each person in the criterion group, and enter his criterion score in the column corresponding to every answer which he has made. Thus, suppose we had individuals whose criterion scores in the test questions were 95, 84, 73, and 65. Suppose that person whose criterion score was 95 answered true to the first question, true to the second, checked the second alternative in the multiple-choice question, and ranked the ranking questions in the order 5, 3, 2, 1, 4. Suppose that the individual

[1] Adapted from McCall, Long, and others, *Teachers College Record*, Vol. 27, pp. 394 ff., January, 1926.

whose criterion score was 84 answered true to the first question, false to the second, checked the third alternative, and ranked the elements in the order 5, 2, 1, 3, 4. Suppose that the individual whose criterion score was 73 answered true to the first question, false to the second, checked the fourth alternative, and ranked the questions in the order 4, 5, 1, 2, 3. Suppose that the individual whose criterion score was 65 answered true to the first question, false to the second, checked alternative 1 in the multiple choice, and ranked the elements in the order 1, 3, 2, 4, 5. They would then be tabulated as in Table III on opposite page.

4. Add each column and divide by the number of entries in the column. This will give the average value of each answer in terms of criterion score. At the same time add the column of criterion scores so that we may find the average of the criterion scores.

5. Subtract the average criterion score in algebraic fashion from the average of each column. The resulting figure will give, if positive, the amount of credit that ought to be given for that answer; if negative, the amount which should be subtracted from the score for any person who makes the particular response.

If the test be scored in this fashion it will give high scores to the kind of people who stood high in the criterion, and it will give low scores to the kind of people who stood low in the criterion. Any elements which were answered in the same fashion by everyone receive zero rating. Any elements which were answered by nobody are eliminated. Any elements which were answered equally often by people who stood high in the scale and people who stood low in the scale receive zero value. The difference between the amount of positive credit which can be given for the right answer to a test question, and the amount of penalty which is given for the poorest answer, is a rough indication of the value of that element.

Sometimes these results are rather surprising. It appears that elements are not always so simple as they seem. For example, in the study of one examination the criterion based on the answers of 2,000 boys showed that the true-false statement, "The Jews persecuted many of the prophets and refused to believe the messages which they brought," was practically valueless. Analysis showed that some boys who stood very high in their religious knowledge knew that this was true. Another group who stood very low in religious knowledge rather

TABLE III. — TYPICAL FORM FOR THE WEIGHTING OF TEST ELEMENTS IN ACCORD WITH CRITERION SCORES

Criterion scores		(True-false)				(Multiple-choice)				IV (RANKING)																						
	Questions	I		II		III				First item					Second item					Third item					Fourth item					Fifth item				
	Responses	T	F	T	F	1	2	3	4	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
95		95		95			95							95			95				95				95				95				95	
84		84			84			84						84		84				84						84						84	
73		73			73				73				73						73	73					73						73		
65		65			65	65				65							65				65					65							65
Etc.																																	
Sum																																	
Average																																	
Subtract Average C. S.																																	
Value																																	

disliked Jews anyhow and marked the statement true as the result of their prejudice. Again it was found that the statement, "It is one of the principal duties of the Christian church to bring about peace and justice on earth," was valueless because everybody answered it as true regardless of his criterion score. One of the very great advantages of this method is that it gives a small, high-class minority preference over the great majority. If a scoring system be based simply on the most frequent answer, it will be very apt not to give due credit to the few who think beyond the majority. Thus, in a multiple-choice question with reference to the best way in which to think about God, it was found that the great majority checked as true the statement, "God should be thought of as all-good, and all-powerful." It appeared, however, that among these particular adolescents those who had done enough thinking so that they placed that in the doubtful column were distinctly superior to the great majority when all their other answers were taken into account. The criterion in this case was based on the results of the total test scored in accordance with the best available judgments. However the test-makers may feel about it, this result, especially if confirmed by a larger group, would indicate that more credit should be given to the person who gives the answer "doubtful." Of course, this might not hold with other groups, and with tests which endeavor to measure traits other than religious thoughtfulness.

The resulting score values may all be multiplied or divided by any figure without injuring the relative values. Likewise if it be desired to eliminate negatives, the same number may be added to each obtained value. The "Brief Test of Religious Education" described on page 72 is scored on the basis of the answers of several thousand boys. The value which each element received was multiplied by such a figure that a perfect score became 100, and the average boy would receive zero. Negative scores on this test mean "below average."

d. Weighting Test Elements

So far as possible, a score should be given for every response which the subject makes. The units of the scoring scheme

should be equal to one another and interchangeable. If two pupils each have 50 points correct it should be a fair assumption that their ability is the same although they may have succeeded and failed at very different points within the scale. This is an assumption which it is very hard to make true. All questions are certainly not of uniform difficulty. There is considerable difference of opinion among test makers as to the value of weighting the elements. If the differences are not very great, and the number of elements is large, it is certain that even complicated weighting systems will yield scores which correlate almost perfectly with the scores reached when no weighting at all is used. It must be remembered in connection with any attempt to weight elements in a test or score card, that the real weight when the score has been totaled is determined not by the absolute size of the credit given to any single item, but by the relative variability among the scores made by different persons on the given item.[1] Thus, if on a relatively unimportant item some people were given zero, others 1, 2, 3, or 4 points, that item might have more influence on the total score than would an item in which the best people receive 20 points, and others 18 or 19, no one receiving less than 18. If weighting is to be surely effective, it should be in terms of the spread of the scores rather than of their absolute size. When the criterion score method of developing the value for each answer is used, the problem of making units equal is fairly well taken care of. Each element is weighted in terms of its differentiating power, and the unit in each case comes from the criterion.

7. Standardization of Tests

The final process in connection with the construction of a test is its standardization. As a result of the suggestions made so far, the test material has been carefully selected, classified, and organized into controlled answer form, the directions have been carefully stated, and the scoring system worked out so that the test will, so far as practicable, measure the thing which it is supposed to measure. Standardization involves four items: objectivity, reliability, validity, and comparability.

[1] See pp. 38–39.

a. Objectivity

An objective test is one that is so constructed that the results will be the same no matter who makes use of it. A manual should be prepared stating the age group for which the test is suitable, the room conditions under which it is to be given, whether tables and desks are to be used, whether an adult supervisor must be present with children, etc. If it is a time test the exact amount of time to be given to each section must be stated. Directions for introducing the test to a group should be suggested, and all possible uniformity in procedure guarded.[1] The scoring system should be stated in such a fashion that any capable clerk can score the paper, reaching the same result which any other capable clerk would reach. One of the best ways of facilitating scoring is to prepare a score sheet which can fit the test exactly, and which shows the amount of credit to be given for each answer at each point. If the test is a completion test, it is necessary to state every answer for which any credit at all will be given, and this requires a wide range of previous experimentation. Scoring is facilitated further by having all the answers in a single column so that one does not have to hunt around the page for them. Frequently, a celluloid or tissue paper score sheet is used so that it can be placed over the answers and any discrepancies stand out immediately. Sometimes rather ingenious devices are used to hasten the scoring; thus, the correct answers expressed in terms of letters may spell a certain word or in terms of numbers they may follow a certain series which can be recognized at a glance. If the scoring system, however, becomes too simplified, it will be recognized by subjects taking the test and they will in some cases make the judgment on the basis of what they believe the scoring system to be rather than on the basis of the test itself.

b. Reliability

The second element in standardization is the determination of reliability. This means that the test must be consistent under

[1] The *Manual* should give names and addresses of author and publisher so that future workers may contribute to test standardization and may obtain more information when needed.

the same conditions, with the same persons it must yield the same results. Ordinarily reliability is expressed in terms of the correlation between the results obtained when the test is given to a group, and when the same test is repeated with the same group, no intervening factors having influenced some persons more than others. It is very difficult, of course, to have the group twice in exactly the same state of mind, health, and happiness. The very fact that they have taken the test once means that some have gained more through the practice than others have gained. Reliability is never perfect under such circumstances.

An ingenious method for determining reliability is now frequently used instead of giving the total test twice. This involves splitting the test into two equivalent sections. Suppose the test consists of 100 elements, we will then choose elements 1, 3, 5, 7, 9, etc. up to 99, and see what the score of each individual would have been had he taken only those 50 elements. Then select the even-numbered elements and see what the score would have been on that test of 50 elements. So for each person taking the test we may find the score on two half-tests taken at practically the same time and under practically the same conditions. The correlation between the scores on one half of the test and the scores on the other half of the test indicates what the reliability of half of the test is. This may be extended by the use of the Spearman-Brown Prophecy formula which states that:

$$r_n = \frac{nr_1}{1 + (n - 1)r_1}$$

where

r_n = the self-correlation of tests which are n times as long as the tests measured in the experiment.

r_1 = the correlation which has been obtained in the experiment.

n = the number of times the tests are to be used, or the number of times as long the desired test is to be.

Thus, for example, suppose we experimentally determine that the correlation between one half of a test and a duplicate half, is 0.63. This is r_1. We wish to find out the self-correlation of the whole test, which would, of course, be twice as long as either half.

That means that n becomes 2. Then r_n, the desired self-correlation. would be:

$$r_n = \frac{2 \times 0.63}{1 + (2 - 1) \times 0.63} = \frac{1.26}{1.63} = 0.77$$

The same formula is used in the same way, to determine what the self-correlation would be if tests were slightly longer or shorter. An experimenter has a group of tests of unequal lengths and reliabilities. He wishes to compare them and to see which one is the most reliable form with which to work. The difference in length means that the longest one has an undue advantage over the shorter ones, in apparent reliability. Would that superiority hold if all were the same length? This he can answer by reducing all tests to one given length, say 100 items. Then for tests that are 200 items long he would substitute 0.5 in the place of n in the above formula, and reduce the apparent correlation to a 100-item basis. For tests that had originally only 75 items, he would substitute 1.33 in the place of n, and see what the self-correlation probably would have been, had he had 100 items instead of only 75. Kelly, Wood, Ruch, and others have checked this formula developed by Spearman and Brown, and usually known as the Spearman-Brown Prophecy formula, and found it to hold very well. It is an empirically determined formula, with no clear theoretical justification. Wood and Ruch both found that it yielded correlations slightly too low when n was high, but the differences were not large or invariable.

The formula may be turned inside out, and used to determine how much longer a test or tests will have to be, in order to yield a given reliability. In this form, (solving for n) it would read:

$$n = \frac{r_n - r_1 r_n}{r_1 - r_1 r_n}$$

A need for this formula arose in connection with the Y M C A examinations. It was found that the average reliability for a single page of their test questions during 1 year, was 0.45. The question was then asked, "How long do the examinations have to be in order to give a good reliability?" For their purposes a

good reliability was defined as 0.80 or above. Then r_n had to be 0.80, and r_1 was known to be 0.45.

$$n = \frac{0.80 - 0.45 \times 0.80}{0.45 - 0.45 \times 0.80} = \frac{0.80 - 0.36}{0.45 - 0.36} = \frac{0.44}{0.09} = 4.9$$

The examinations would have to be 5 pages long, or practically so, to attain a reliability of 0.80. In actual practice this meant reducing type so as to get five times as much material into a 4-page folder as had once been on a 1-page folder, or else using an 8-page folder which would cost much more for preparation, printing, distribution, and scoring, but which would yield only slightly higher reliability, probably (using the first formula with $n = 8$) about 0.87. The actual procedure was to select a type of questions which yielded a higher reliability than did the ones previously used (completion and rank questions had been found to give highest reliabilities) and to advise the use of a 4-page test.

Reliability increases directly with the length of the test and with the variability among the group to whom it is given. A test may appear very unreliable when given to a group of 200 fifth graders all of whom have white, native-born, Protestant, middle-class parents, and yet appear very reliable if given to 200 pupils who vary over a wide range, some of them representing very high intelligence, others being at the feeble-minded level, some of them having had excellent ethical training, others having grown up on the street. When the reliability of a test is stated, therefore, there should be stated with it some measure of the number and variety of persons in the group upon whom reliability was determined. Usually a brief description of the group should be given and a statement of the standard deviation of the test scores within that group. Suppose an individual uses a test in his community, and finds that the standard deviation on the test for his community is approximately only half as much as it was in the group upon whom the author of the test standardized it. He may well be assured that the reliability is less in his situation. He will probably want to use not one test, but several. Reliability coefficients taken by themselves may be misleading. A

test with a reported reliability of 0.90 may be inferior to one with a reliability of 0.70 obtained in a more homogeneous group.

c. Validity

The third feature to be studied in connection with standardization is the validity of the test. This answers the question, "Does the test measure what it purports to measure?" If the scoring has been developed upon a criterion basis, the probability is very good that the test's validity is acceptable. At least the test will be as valid as the criterion upon which it was developed. It will usually prove wise, however, to repeat the process with a second or third group. Lentz found that after criteria had been set up within one group, and test elements selected which separated out the good from the bad in that group, when he came to apply these to a different group they did not hold very well. Some elements which had seemed to be rather useless before now stood out, some elements which seemed to be particularly good before, reversed themselves in the new situation. Some that had previously been answered only by the very best appeared in the second case to be responses which were made only by the very poorest. Such extreme results will, fortunately, not always arise to dismay the investigator. It is probably safe to state, however, that a test, the scoring of which has been developed upon a limited group of persons who have been intensively studied, should be applied to a few other groups which are studied with the same care. If it is found that the test differentiates the good from the bad in a number of groups, it may well be regarded as a valid test for differentiating the particular thing that it is supposed to distinguish.

The validity of a test may be further studied by selecting new criteria. If the judgments of teachers have been used previously, suppose the judgments of parents be substituted when it comes to making a further validation. The larger the number of criteria against which the test seems to check satisfactorily, the surer is the evidence for its validity.

Five methods for studying validity were discussed in connection with the general problem of measurement and rating, on pages 47–53. First, tests may be validated by checking them

against other and supposedly better measurements, such as previously standardized tests, long batteries of tests, and careful ratings of small groups. Second, tests may sometimes be validated by using objective evidence about persons, indicating how far the test was really right. Third, groups may be used which are distinctly different from one another with reference to the trait being tested, and the test may be validated by the difference it reveals between such groups. Fourth, validation is often helped by studies which show how far the test measures other things than the one it is supposed to measure, such unwanted things, for example, as home environment, schooling, or reading rate. Fifth, it may be shown that a certain test is valid because of the changes in score which take place after people have gone to certain conferences, become converted, studied certain books, acquired a fortune, or been subjected to similar stimuli which are supposed to have brought about changes in the particular thing being tested. Finally, it was pointed out that in many cases, reliability is the only important question. Given the reliable test, it can later be found how well that test will predict good citizenship, or miserliness, or ability to become a recognized leader. It is unnecessary, so some would say, to match up tests with words which describe imaginary abstractions called traits and characteristics. Let the test be judged by its accomplishments, without pretending that it is a test of this or that.

d. Comparability

A fourth step in standardization is the establishment of norms so that new scores can be compared with certain standards. For educational tests this has been simple. Tests were given to 1,000 pupils in the fourth grade, 1,000 in the fifth, in sixth, etc. The average and the standard deviation for each group were reported, and any new class could tell immediately where it stood with reference to the country as a whole, in so far as those tested were representative of the country as a whole. In religious education, as has been noted, age and grade norms seem to be of much less significance. Hartshorne and May have found a zero correlation between mental age and certain ethical conduct. How then shall we report scores in such a way that they will be of

maximum help to other workers? At least we can report the scores for the average pupil in each type of church school, public school, etc. Probably norms in terms of intelligence and home background would be the most useful available at present. The maker of a test of religious attitudes would state that the average score for people with an intelligence quotient between 90 and 100 who come from a home background rated in the upper quartile on a certain scale was 42, whereas with increasing intelligence the scores increased and with poorer home backgrounds the scores were correspondingly lower. Religious education is hardly yet at the point which makes it possible to prescribe the form in which test scores shall be reported. About all that can be said at the present time is that for every published test there should be available a statement of the groups upon which this has been tried and the records which these groups made. Thus, for the test of fair-mindedness, called a "Survey of Public Opinion on Social, Religious, and Economic Questions,"[1] the norms have been reported in terms of normal school students, college students, persons chosen as most prejudiced, persons chosen as most fair-minded, Methodist ministers, students in graduate theological schools, etc. For interest analysis blanks, profiles are frequently given for persons going into professions as contrasted with persons going into trade or industry. Whatever basis seems to offer light by which any person who takes the test in the future may make his score more meaningful should be published under the general heading of norms.

8. Criteria for a Good Test

Whether the test be one which is ordered from a publisher or one which is constructed by the experimenter, the following criteria may be found useful in judging its value and suggesting points at which it could be improved.

a. The material of the test should be such as to have a good effect upon the subject at the time when he takes it.

b. The material of the test should be such as to have a good effect on the procedure of teachers, religious educators, and

[1] See Watson, "The Measurement of Fair-mindedness," Teachers College Bureau of Publications, New York, 1925.

others who may use the test in evaluating progress. Public education affords a caution well illustrated by the early use of standard tests in the fundamental operations in arithmetic. Those tests emphasized certain phases of arithmetic. It was soon found that teachers, in order to make a good showing for their class, were stressing items found in the test in the precise form in which they were found in the test. Sometimes the items were not particularly important for pupils' working knowledge of arithmetic in life, and sometimes the form was a distinct handicap. Indeed, let us hope that the tests set forth by religious educators will not only be free from direct encouragement to poor teaching, but will offer distinct incentive for improving curriculum and method. This is not to be interpreted as meaning that the tests should be curricula. The statement is often made to test makers, "Test, don't teach." It is true that the function of measurement is not the same as the function of instruction, but measurement will certainly affect instruction, and that effect should be appraised.

c. The test should be interesting. Most of the tests in moral and religious education do seem to prove interesting to people who take them. The concomitant attitudes are certainly more rewarding for a test of this sort. Other things being equal, variety adds to the interest of a test, but probably the more fundamental factor in making a test interesting is that it deals with matters which are closely related to the major emotional centers in the lives of the people who take it.

d. The test should be objective so that no matter who scores it, the score will be constant.

e. The directions should be carefully, completely, and clearly stated.

f. The test should embody some very easy questions and some which are so hard that no one can answer all of them. It should contain no elements which everyone will get correct, and no elements which everyone will find too difficult.

g. The test should be comprehensive; that is, the material should be selected to cover the entire field which is being tested. Most attention should be given to those parts of the field which

are most important. Ordinarily the material should be checked against some criterion for comprehensiveness. Comparatively few ethical tests have thus far made use of any exhaustive scheme for classifying human conduct such as is being brought together by the Character Education Institution[1] in Washington. Where the tests are supposed to cover a certain book, such as the Bible, it is somewhat easier to build a criterion for comprehensiveness. Even there the problem of the relative emphasis to be given, for example, to Micah and Noah, has not been solved to the satisfaction of most teachers of religious education.

h. The test should involve as little writing as is practical. Where a check mark is adequate, it should be used. It takes more time and effort to prepare a set of alternatives which will be so complete that no writing will be necessary, but in the long run, if the test is widely used, much time will have been saved.

i. The test should contain no catch or trick questions.

j. The test should measure the one thing which it sets out to measure, and avoid the lure of interesting bypaths. There are some tests being made which afford an opportunity for the subject to register all sorts of interesting points of view. Neither the subject who takes the test nor the person who made the test has much idea what the statement is good for after it has been made. The test should be pruned of the many interesting but irrelevant items which could so easily be connected with it.

k. Each element in the test should be independent of the answer to every other element. It is distinctly unfair to have the answer to one question depend upon having reached the right answer in the previous one. For the purpose of study on elements and the selection of the best ones, the test elements should be so complete that they can be moved around in a test without destroying their meaning. Incidentally it should be noted that the position of an element in the test has a distinct effect on its reliability and validity. Items must not be juggled in order, if the results are to be comparable with previous administrations.

[1] Chevy Chase, Washington, D. C., Milton Fairchild, Director.

l. The test should require as little reading as possible. If concrete cases are described, great care should be used in boiling down the description to the few essential elements. It has been estimated that for some tests which are supposed to be tests of ethical ability and which occupy 40 minutes of time, 70 per cent of that time is spent in actual reading, with no allowance for weighing, evaluating, choosing, deciding. Differences in reading ability among subjects are very great. In the ordinary classroom the best pupil not infrequently reads three times as fast as the poorest. When the attempt is to measure something besides reading ability, this factor may play havoc in the apparent results.

m. The score should be numerical and should be as meaningful as it is possible to make it.

n. The test should be non-fakable. Pupils should not be able to raise their score at will. Likewise the test, where possible, should be non-coachable; that is, it should not be possible for pupils who have been let in on the trick to make an undeservedly high score.

o. The elements within the test may well be arranged in order of difficulty, the easy ones coming first.

p. Where possible the score should be diagnostic or at least analytical; that is, it should not only classify a pupil as good or poor, but it should indicate much more specifically wherein and why he is good or poor.

q. The reliability of the test should be established. This, of course, is directly dependent on the length of the test, and the variability within the group tested. For some purposes such as the obtaining of group average, a brief test with relatively low reliability is sufficient. Reliabilities of 0.40, 0.50, or 0.60 may be used if it is not intended that any dependence should be placed upon an individual score. The better standard tests in education have reliabilities of 0.80 and above. Some educators believe that scientific measurement is quite impossible with instruments having a reliability below 0.95.

r. Such evidence as it is possible to obtain should be offered for the validity of the test. It should rarely be necessary for a test to be accepted as valid merely upon prima facie evidence. It may or may not really measure that which it appears to measure.

EXERCISES

1. Construct a sample test question, bearing upon the study of the Gospel according to Mark, falling in each of the following types:

1. Essay type.
2. True-false.
3. Other two-answer.
4. Completion.
5. Word-phrase answer.
6. Multiple choice.
7. Degree of truth.
8. Ranking.
9. Matching or pairing.

2. Try these tests out, by passing them on to some other member of the class to answer. Were any ambiguities found?

3. Which of the questions prepared most effectively gets beyond simple information to the questions of understanding, thoughtfulness, and change of purpose which might grow out of such a study?

4. Construct a sample test question illustrating each of the following difficulties:

1. Directions not clear to a person never having taken such tests before.
2. Double or triple negative in the "false" statement.
3. Ambiguous completion because too much left out.
4. Obvious completion, anyone could answer it.
5. Multiple choice with only one real possibility.
6. Multiple choice with alternatives all poor.
7. Multiple choice with too many good alternatives.

5. Prepare a new-type test which will measure mastery of the material contained in Chapter V. Let the teacher select the best of these questions and administer them to the entire class. Then appoint committees or otherwise assign the following tasks:

1. What was the difficulty of each question element in terms of the per cent of pupils failing to answer it?
2. What was the percentile score of each member of the class? Translate these to S. D. units, using Table XXXIII.
3. Using the obtained score of each class member as his criterion score, find the value to be allotted to each possible answer to each question element.
4. Split the test elements into two halves and indicate roughly the reliability by showing the rank of each pupil in the class on each half of the test. If the test is reliable, those ranking high in one half will rank high in the other, and *vice versa*.

6. Obtain one of the published tests listed in Chapter IV, and score it on the basis of 10 possible points for meeting each criterion suggested in this chapter. A perfect score would be 180.

CHAPTER VI

STATISTICAL METHODS

1. The Service of Statistics

Statistics seem to divide people very sharply into two groups; some like them, others decidedly do not. Many adult attitudes are probably due to experiences with arithmetic in grade school and algebra in high school. Some dislike may be due to never having had the experience of insight which mathematics may bring. Pure mathematics, like pure music, has its romance of form and relationship. It may bring with it the sense of the achievement of perfection and harmony, the thrill of thinking in the terms in which the Maker of the Universe must have thought.

Applied mathematics has been more than justified in every realm of science. Until it became possible to relate observations in mathematical terms, knowledge remained on the level of a collection of experiences. The collection, of course, may be valuable, but the parts seem unrelated, the total mass of data is confused and incoherent. Sometimes a genius is able to gather together such data within the scope of his tremendous comprehension, and to see the truth which lies hidden in the confusion. Statistical processes make it possible for the ordinary individual to discover true meanings which only the genius would be able to see without the aid of statistical tools. Anyone can illustrate this by trying to compare two series of scores before they are arranged in any order or have been averaged or graphed. Statistical processes telescope many observations into a few terms which embody the whole but which can be more easily grasped and compared. They are not designed to be confusing. By the use of statistical tools we can compare groups with reference to their average level of achievement, and with reference to their

distribution, spread, variation, or homogeneity. We can discover the causal factors which any two traits or results have in common; we can predict from certain data what the result will be under very different circumstances, and we can tell the error of the prediction. We can jump from the finite to the infinite, predicting what the results would have been had we had inconceivably large numbers of cases.

a. Limitations of This Presentation

It is clearly impossible to present a fair discussion of statistics in a single chapter.[1]

The presentation here has three aims:

1. It is hoped that readers may acquire the ability to interpret statistics which they find in reports of experiments. For this purpose the sections in smaller print should be omitted. Read only the general discussions of usefulness and meaning of each measure.

2. Teachers of classes may wish to train students in the actual computation and application of these techniques. The statements of procedure and illustrations here given form a basis for such teaching. It is assumed that no one will attempt to teach a class statistics merely on the basis of the material in this chapter. Because of the brief and condensed form, many ex-

[1] The following texts will be found very useful for supplementary work. In general, more complete and more advanced works are placed later in the list.

Rugg, "A Primer of Graphics and Statistics," Houghton, Mifflin Co., Boston, 1925.

Otis, "Statistical Method in Educational Measurement," The World Book Co., Yonkers, 1925.

Garrett, "Statistics in Psychology and Education," Longmans, Green, and Co., New York, 1926.

Thurston, "The Fundamentals of Statistics," The Macmillan Co., New York, 1925.

Thorndike, "An Introduction to the Theory of Mental and Social Measurements," Teachers College Bureau of Publications, New York, 1913.

Yule, "An Introduction to the Theory of Statistics" (Fifth Edition), Charles Griffin and Co., London, 1919.

Kelley, "Statistical Method," The Macmillan Co., New York, 1923.

Rietz, "Handbook of Mathematical Statistics," Houghton, Mifflin Co., Boston, 1924.

Whittaker and Robinson, "The Calculus of Observations," Blackie and Son, Ltd., London, 1924.

planations, assumptions, and limitations have been omitted. These the teachers should be in a position to supply.

3. Many workers in religious education and allied fields have had statistical training at some previous time, but wish to utilize this chapter as a condensed working handbook.

All of the formulae which will be presented in the following pages are expressions of relationships among measures. These might be measures of distance, weight, monetary value, death rate, intelligence, knowledge, or agreement with an ethical standard. Like other relationships in pure mathematics, they are quite as useful to the astronomer, the physicist, the biologist, the biometrician, the actuary, the psychologist, or educator, as to the religious educator. The assumptions upon which they have been derived are simply refined common sense. These assumptions vary little from one sort of measure to another. Consider, for example, the foremost element in these calculations, the curve of normal distribution[1] which occurs in any realm where purely chance factors are influencing a series of measurements. If 20 pennies be tossed in the air 1,000 times, purely chance factors will tend to influence the number of heads and tails in each toss-up of the 20. It will appear, therefore, that a great many times there were 10 and 10. Almost as many times there were 9 and 11. The occurrence of 2 heads and 18 tails, or of 1 head and 19 tails, would be relatively very rare. A similar distribution may be found with many human measurements. So many causes influence the height of the high school pupils in New York that if they were stood in line there would be relatively few very short ones or very tall ones, but many clustering around the average. If 10,000 observers were to watch a child at play and rate him on cheerfulness, a similar phenomenon might be observed. Some would rate him low, others exceedingly high, while most of the group would accumulate near the center of the total range of measurements. Of all the people in churches last Sunday, it is probable that only a few were tremendously impressed, and only a few utterly unaffected. In between, near the average degree of impression, would be the great mass of

[1] See illustration, p. 183, and discussion on p. 181.

people. In any and all of these cases, accurate recording of very large numbers of cases shows that they tend to fall into a given mathematical form. The curve which would be formed if the measures were graphed has an equation capable of theoretical and practical verification. It becomes convenient, therefore, to know how to find the middle of this curve and to measure distances out from it. It helps us to know from a relatively few cases about what we might expect if we had many more. It gives us a basis upon which we can compare standing in one measurement with standing in a different one.

It may seem that amazingly complicated formulae have sometimes grown out of this attempt to use data fruitfully. The person with little training in mathematics may eye with suspicion the application of square roots to ideas of God. Perhaps it may help him to remember that it is the measurement which is being thus added to others or squared or divided, not the idea itself. The measurement may be good, bad, or indifferent. At best it will surely be difficult and somewhat inaccurate. But since it is a measurement it is subject to the thousands of unnamed and indescribable errors and variations which we call "chance." A different group, with other persons, would surely have yielded other measures. The measurement as a measurement may be treated, then, just as any measurement in other scientific fields is treated. The length of a rod, or the strength of a belief, or the length of a given person's life are all matters of probability, measures differing in degree of certainty with which they can be approximated, not different in nature.

Authority for the particular methods of treatment here recommended must be sought in the field of the theory of statistics. The reader anxious to understand the assumptions and implications of the mathematics involved may well begin with Thorndike's "Mental and Social Measurements," and proceed through Yule and Rietz (see note, p. 172). He will need, of course, the tools of differential and integral calculus. Other students interested only in a practical understanding may wish to accept the formulae on the simple fact that they are not unusual and intricate compilations, but are tools which are now almost universally

used in educational research. They are no more likely to be questioned when used for the purpose here indicated and within the limits of the interpretation here suggested, than would be the use of a per cent in reporting increase in membership, or an index figure in reporting economic facts. In any case when reading a section for the first time it may prove helpful to read the "purpose" and "interpretation" of each measure before trying to master the technique of computation.

In spite of the limitations of space this chapter will attempt to go a little further into the applications of correlation, partial correlation, regression, and multiple correlation than is usual with elementary textbooks. This is done because it is difficult for a student to see much value in the simpler statistical measures. The service of statistics in revealing hidden relationships stands out most clearly in some of the more advanced processes. From the standpoint of appreciation, if not of computation, it seems desirable to include problems of a somewhat advanced sort.

Only one method for computing each statistical measure will usually be suggested. The method chosen is the result of a process of elimination. Three classes of theological students have, at different times, been experimented upon to find out which of the various methods of calculating correlation, standard deviation, etc., they found most congenial. They did not always choose the method which is most revealing to the statistician. The methods of computation presented here are the results of their choice. They involve only processes which any individual who has studied elementary algebra should be able to carry through successfully.

A serious problem in connection with the teaching of statistics to those who wish to be experimenters in the realm of character education and religious leadership lies in the dearth of applications which have previously been made. Religious education has, as yet, few critical and exhaustive statistical analyses of curricula, methods, and results. Even were such illustrations available, however, it is quite probable that students would still have to learn statistics more by their own activity than by study of the results of others' experiments. It is frequently stated

that no one ever learns statistics until he tries to work out an original problem. In large measure this is true. There are few subjects in which the gap between understanding an explanation in a book and being able to make use of processes in actual experimental situations is wider. It is, therefore, recommended that classes be divided into small groups, each of which secures some live and interesting data and follows through the discussions in statistics, making constant application to this experimental material.

For purpose of illustration, a class of 25 pupils has been chosen and assigned a series of hypothetical scores. It is assumed that pupils in this group have taken an intelligence test, a test of religious attitude and information, and a test of ethical knowledge. So far as practicable, each technique presented is illustrated in terms of these three sets of scores for the class of 25 people.

2. Distributions of Scores. *a. Order Distribution*

The first problem which faces an experimenter who has given some test or collected certain observations is the tabulation and arrangement of the material. Several forms of distribution for scores are commonly employed. The simplest, perhaps, is to list the names of the individuals in alphabetical order, placing after each the score obtained. This has been done with the intelligence scores and the ethical scores in Tables XII–XVII. Another frequent practice is to make an order distribution in which the highest score is placed at the head of the list, then the next highest and so on down to the lowest score. This is illustrated in Table IV. Such an arrangement is usually needed in addition to the alphabetical distribution because while it is more difficult to locate any single pupil, it is much easier to note the range of class scores, the approximate middle of the group, and the relative position of any individual.

b. Rank Distribution

In order to make the relative position clearer pupils are sometimes listed in rank distribution. A rank distribution assigns 1 to the best person, 2 to the next, 3 next, and so on. This is illustrated in Table V. Thus, for the intelligence test, John

TABLE IV. — ORDER DISTRIBUTIONS

Intelligence		Ethical score		Religious score	
Pupil	Score	Pupil	Score	Pupil	Score
John	146	James	28	Selma	98
Selma	142	John	27	Margaret	95
Harold	140	Theron	27	Fred	94
Fred	140	Clarence	27	Marion	90
Theron	137	Selma	26	Gordon	86
James	132	Dan	26	Florence	81
Margaret	131	Harold	26	Theron	77
Clarence	131	Jessie	26	Dorothy	75
Marion	129	Jeanne	25	Walter	74
Helen	128	Donald	25	John	74
Jeanne	128	Margaret	25	Howard	73
Marvin	128	Ben	25	Dan	71
Dan	127	Gordon	24	Clarence	67
Ben	127	Marvin	24	Marvin	62
Gordon	127	Fred	23	Ben	58
Samuel	126	Tony	23	Samuel	54
Jessie	126	Marion	22	Harold	53
Donald	126	Helen	21	James	50
Florence	126	Florence	20	Jeanne	45
Walter	125	Howard	20	Helen	41
Dorothy	125	Walter	19	Donald	40
Howard	124	Dorothy	18	Jessie	37
Sarah	122	Sarah	17	Oscar	32
Oscar	117	Oscar	16	Sarah	29
Tony	105	Samuel	15	Tony	19
Total	3,215	Total	575	Total	1,575
Mean	128.6	Mean	23.0	Mean	63.0
Median	127	Median	24	Median	67
Mode	126	Mode	25 and 26	Mode	74
Q_1	125	Q_1	20	Q_1	42
Q_3	131	Q_3	26	Q_3	74
$Q = \frac{131 - 125}{2}$		$Q = \frac{26 - 20}{2}$		$Q = \frac{74 - 42}{2}$	
$= 3$		$= 3.0$		$= 16.0$	
Range $= 146 - 105$		Range $= 28 - 15$		Range $= 98 - 19$	
$= 41$		$= 13$		$= 79$	

secured the highest score, Samuel was second, and so on. We note that Harold and Fred made the same score. It is obviously unfair to give one more credit than the other receives, so the ranks which they would have received, 3 and 4, respectively, are averaged, and each one is given a rank of $3\frac{1}{2}$. Note that Theron's rank then becomes 5 because there are four superior to him.

An easy rule-of-thumb for making out a rank distribution is to number the pupils in order as has been done at the left of the names in Table V, and then to give every pupil a rank corresponding to the number which stands before his name, except in the case of pupils who are tied. In such cases, the numbers standing in front of the names should be averaged and each of the pupils given the average value. Thus, in the case of the ethical scores, John, Theron, and Clarence have all received the same actual score. According to the listing, one is second, one is third, and one fourth. Adding 2, 3, and 4, we have 9, and dividing by 3 we discover that the rank to be assigned to each one should be 3. If it is desired to check a rank distribution, the ranks assigned may be added. The sum should be equal to the sum of the numbers in front of the names.

c. Frequency Distribution

When a large number of scores have been obtained, it is frequently desirable to group them. A frequency distribution tells how many persons made each score, or how many made scores falling within a certain range, such as 90–94, 95–99, etc. The frequency distributions illustrated in Table VI represent common practice with reference to grouping scores.

In the case of the intelligence scores, a step interval of four has been chosen and the number of pupils making scores between 145 and 148 inclusive is shown in the first line to be 1. The number making scores between 141 and 144 is also 1. Three made scores between 137 and 140. The total number in the frequency column should, of course, correspond to the number of scores obtained. In similar fashion, the scores from the ethical test and the religious test are reported. In the case of the ethical test, the scores were so close together than a step interval of one was used.

d. Graphic Distribution

The frequency distributions lend themselves very readily to graphic presentation. A graphic distribution is a diagram showing how many persons made each score. Table VII shows

how the distribution of intelligence, ethical, and religious scores looks when it is plotted out on squared paper rather than simply stated in figures. Step intervals used are the same as in the frequency distributions. In the case of the intelligence scores, a cross has been placed in a square for every score, then the tops of the columns have been joined by a heavy black line which gives a general picture of the distribution. It shows that there was one score distinctly removed from the rest, and that by far the largest number of people made scores between 125 and 128.

TABLE V. — RANK DISTRIBUTIONS

Intelligence		Ethical score		Religious score	
Pupil	Rank	Pupil	Rank	Pupil	Rank
1. John	1	1. James	1	1. Selma	1
2. Selma	2	2. John	3	2. Margaret	2
3. Harold	3½	3. Theron	3	3. Fred	3
4. Fred	3½	4. Clarence	3	4. Marion	4
5. Theron	5	5. Selma	6½	5. Gordon	5
6. James	6	6. Dan	6½	6. Florence	6
7. Margaret	7½	7. Harold	6½	7. Theron	7
8. Clarence	7½	8. Jessie	6½	8. Dorothy	8
9. Marion	9	9. Jeanne	10½	9. Walter	9½
10. Helen	11	10. Donald	10½	10. John	9½
11. Jeanne	11	11. Margaret	10½	11. Howard	11
12. Marvin	11	12. Ben	10½	12. Dan	12
13. Dan	14	13. Gordon	13½	13. Clarence	13
14. Ben	14	14. Marvin	13½	14. Marvin	14
15. Gordon	14	15. Fred	15½	15. Ben	15
16. Samuel	17½	16. Tony	15½	16. Samuel	16
17. Jessie	17½	17. Marion	17	17. Harold	17
18. Donald	17½	18. Helen	18	18. James	18
19. Florence	17½	19. Florence	19½	19. Jeanne	19
20. Walter	20½	20. Howard	19½	20. Helen	20
21. Dorothy	20½	21. Walter	21	21. Donald	21
22. Howard	22	22. Dorothy	22	22. Jessie	22
23. Sarah	23	23. Sarah	23	23. Oscar	23
24. Oscar	24	24. Oscar	24	24. Sarah	24
25. Tony	25	25. Samuel	25	25. Tony	25
Totals	325		325		325

TABLE VI. — FREQUENCY DISTRIBUTIONS

Intelligence (Step interval 4)		Ethical score (Step interval 1)		Religious score (Step interval 10)	
Score	f	Score	f	Score	f
145–148	1	28	1	90–99	4
141–144	1	27	3	80–89	2
137–140	3	26	4	70–79	6
133–136	0	25	4	60–69	2
129–132	4	24	2	50–59	4
125–128	12	23	2	40–49	3
121–124	2	22	1	30–39	2
117–120	1	21	1	20–29	1
113–116	0	20	2	10–19	1
109–112	0	19	1		—
105–108	1	18	1	N	25
	—	17	1		
N	25	16	1		
		15	1		
			—		
		N	25		
$Median = 125 + \frac{8.5 \times 4}{12}$		$Median = 24 + \frac{1.5 \times 1}{2}$		$Median = 60 + \frac{1.5 \times 10}{2}$	
$= 125 + 2.8 = 127.8$		$= 24 + .75 = 24.8$		$= 67.5$	
$Q_1 = 125 + \frac{2.25 \times 4}{12}$		$Q_1 = 20 + \frac{1.25 \times 1}{2}$		$Q_1 = 40 + \frac{2.25 \times 10}{3}$	
$= 125.8$		$= 20.6$		$= 47.5$	
$Q_3 = 129 + \frac{2.75 \times 4}{4}$		$Q_3 = 26 + \frac{1.75 \times 1}{4}$		$Q_3 = 70 + \frac{5.75 \times 10}{6}$	
$= 131.8$		$= 26.4$		$= 80.0$	
$Q = \frac{131.8 - 125.8}{2}$		$Q = \frac{26.4 - 20.6}{2}$		$Q = \frac{80.0 - 47.5}{2}$	
$= 3.0$		$= 2.9$		$= 16.3$	

For discussion of the meaning of the formulae here illustrated, see pp. 184 and 187.

In the case of the ethical scores, a slightly different procedure has been illustrated. As before, a cross was made to represent each pupil's score, but rather than using a blocked line for each

column, a smooth curve has been drawn indicating what the distribution would probably look like if there had been many more cases, and the divisions used had been finer. It shows a distribution with most of the scores piled up near the high end of the scale, and a few stringing down at the low end. After the same test had been given to 10,000 unselected pupils, the smooth line joining the tops would have much more nearly resembled Table VIII, the so-called "curve of normal distribution."

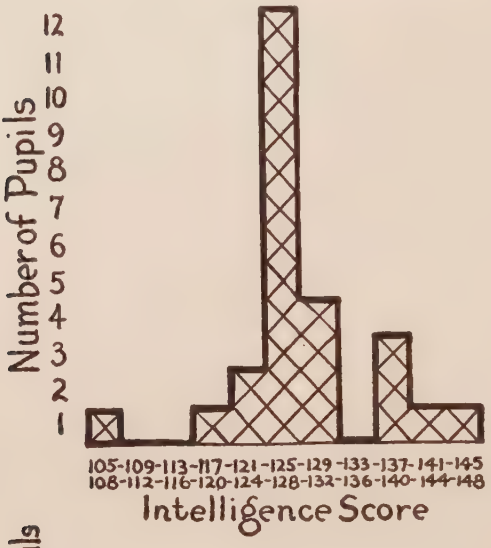
A curve like that obtained from the ethical scores is said to be skewed. When it is skewed toward the high end, it is called a negative skew. This may mean that the test was too easy for the pupils, or that if the test was normal, perhaps the group contained many very good ones and only a few poorer ones. Again it may be interpreted as meaning that the real average for the test belongs at 26 or 27, and that these pupils were the low half of a normal group. In the case of the religious scores a form of graphic presentation is illustrated which may be slightly clearer for purposes of publishing results for people not interested in statistics. It symbolizes rows of persons standing at each score point, and makes it clear that a great many more obtained scores between 70 and 80 than obtained scores at any other point on the scale.

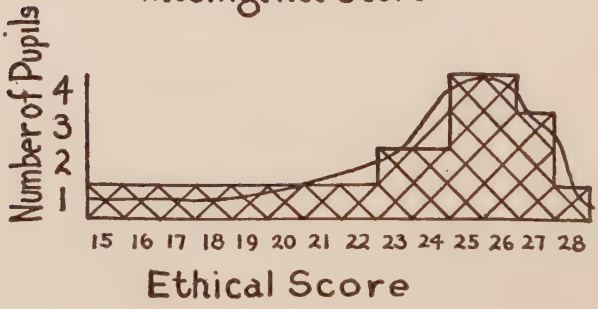
Sometimes circles, squares, dots, bars, or other marks are used to indicate the distribution of scores. Sometimes one graph may be drawn on the same sheet and scale with another, and so shaded that the differences between two or more groups stand out clearly.

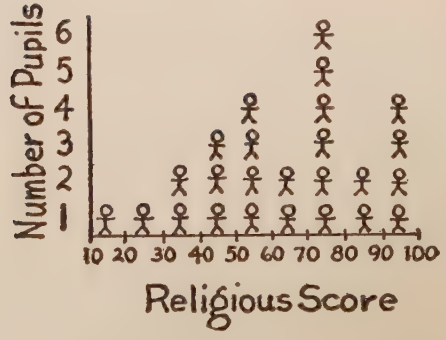
The normal curve outlined in Table VIII indicates what the distribution of any extremely large series of measurements of a single characteristic would be. It has been found to correspond very closely to the distribution of many human traits.[1] If we were to measure the weight of 10,000 people we will find a few at the light end and a few at the heavy end with the great majority standing near the middle. The same might be expected to be true if we measured length of finger, or speed of pulse, or speed of

[1] Some characteristics, such as eye color, sex, tendency to have twins, specific deformities, etc., tend to fall in distinct classes, rather than in a normal distribution.

TABLE VII. — GRAPH DISTRIBUTIONS

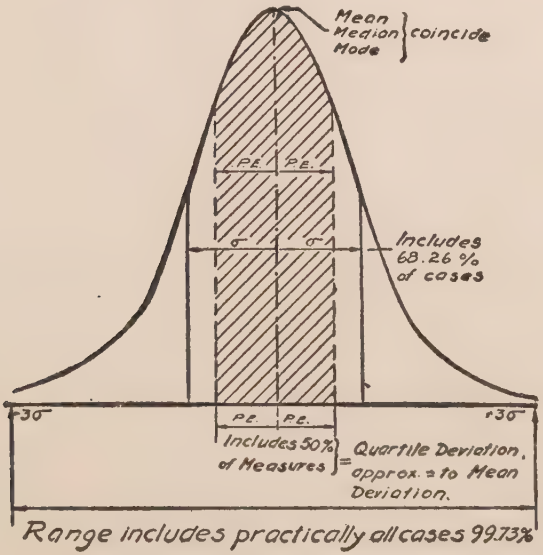






reading, or the kind of intelligence which intelligence tests measure, or degree of beauty. Many statistical processes are based upon the assumption that the data more nearly resemble this normal curve than any other type of distribution. Of course, this does not always hold true, and allowances must be

TABLE VIII. — A CURVE OF NORMAL DISTRIBUTION



made when groups are small and when the graphic distribution does not correspond very closely to the normal curve.[1]

3. MEASURES OF CENTRAL TENDENCY

There are three commonly used measures of central tendency, the mean or average, the median or central score, and the mode or most frequent score.

a. The Mean, or Average

Purpose: The purpose of the mean or average is to aid in the

[1] Theoretically, the normal curve extends out to infinity in both directions, the curve approaching nearer and nearer to zero. Actually the amount lying beyond a distance of five standard deviations (see p. 193) each side of the center is so small that it is not customary to take into account that part of the curve which lies beyond the central 10 S. D. units. Not one case in 10,000 would fall beyond these limits.

comparison of one group with another. High scores tend to be balanced by low scores, and the average usually falls near the center of the group.

Computation: The mean or average is computed in the usual fashion by simply adding the scores and dividing by the total number. Thus, in Table IV we find that the total of all the intelligence scores was 3,215 which, when divided by 25, gives 128.6, the mean or average of that distribution. In like fashion the mean for the ethical scores has been found to be 23, and that for the religious scores 63.

Interpretation: The mean or average is the most commonly used measure of central tendency, and requires little interpretation. When we say that Adventists are more generous than Episcopalians, we mean that the *average* Adventist is more generous than the *average* Episcopalian. It does not, of course, hold true for every person within either group. It should be noted also that it is easily influenced by extremely high or extremely low scores. The score of Tony, for instance, on the intelligence test was 12 points below that of the next lowest. If he had not been included, the mean would have been 130 instead of 128.6. It is possible to have an average with practically every score in the group falling above it (or below it), so great may be the influence of one or two extremes. In the normal distribution the mean coincides with the median and the mode

b. The Median

Purpose: It is sometimes desired to find a measure of central tendency which is not so much affected by extreme scores; one which shows where the real center of the group falls without regard to certain individuals who may have been distinctly different from most of the people in the group. The median also serves the purpose of those who wish to find an average very quickly when there are a great many scores with which to deal. The finding of an average, even with an adding machine, is rather laborious, whereas if an order distribution or a frequency distribution is at hand, the median can be found very quickly.

Computation: Two methods for finding the median are suggested. First, if the scores are arranged in order distribution and there are n scores

then divide $(n + 1)$ by 2 and count up to the score indicated.[1] In the examples given in Table IV, there are 25 scores. To find the median we will add 1 to 25, giving us 26, and divide by 2 giving us 13. The thirteenth score, beginning at either end, gives us 127 for the intelligence test, 24 for the ethical score, and 67 for the religious score. In this case it is clear that it makes no difference whether Tony's score be 105 or 25. He simply counts as one more individual below this middle point.

The median may also be calculated from a frequency distribution such as is illustrated in Table VI.

To find the Median (M) from a frequency distribution:

1. Find half the number of cases (n).
2. Count up, from the bottom, until a step interval is reached which will include all of the cases included in the half which is being counted.
3. Let B stand for the beginning of that step interval, within which the median falls.

Let r stand for the remainder yet to be included to make a complete half of the number of cases.

Let f stand for the number of cases in the f column, for this interval.

Let S stand for the size of the step interval.

Then $M = B + \frac{r \times S}{f}$

Thus, in the case of the intelligence scores, half of the 25 cases would be 12.5. Starting at the bottom, we find that the first step interval uses up 1 of these $12\frac{1}{2}$ cases, leaving us a remainder of $11\frac{1}{2}$. The next two intervals have a frequency of zero each; so they do not help use up the remainder of $11\frac{1}{2}$. The interval 117–120 uses up another 1, leaving $10\frac{1}{2}$, the next interval uses 2, leaving $8\frac{1}{2}$. The interval 125–128 would more than use up the remainder of $8\frac{1}{2}$; so somewhere in that interval falls the median. B is thus 125, r is $8\frac{1}{2}$, the f for that interval is 12 and the step interval, as noted at the top of the column, is 4.

Again, using the ethical score, we find that counting up $12\frac{1}{2}$ cases brings us into the interval marked 24, with 11 cases previously checked off, but with a remainder of $1\frac{1}{2}$. The formula as stated then begins with B (24) and above the line we have r (1.5) and S which is just one for this table. The f for the step interval 24 is 2. The example is worked out as is the first one, in Table VI.

Interpretation: Medians, like averages, afford a basis for comparing one group with another, a basis which is quickly computed and unaffected by extreme scores. The median should

[1] More exactly, this is known as the midscore rather than as the true median. In many test manuals, however, this midscore is called a median. It is easily found, and does not differ markedly from the true median in size or significance.

be used rather than the average, if a few very high scores or a few very low scores are likely to have an undesirable influence in weighting the outcome. When groups are large, comparison on the basis of medians gives almost the same results as would comparison on the basis of means.

c. The Mode

Purpose: The purpose of the mode, as its name connotes, is to indicate the custom or style or popular score so far as the test is concerned. Speaking more carefully, it is the most frequent score.

Computation: Thus, in Table IV it appears that for the intelligence tests, the mode is 126, that score appearing four times whereas no other score appears more than three times. In the religious scores, the mode is 74, that score appearing twice. In the ethical scores there are two modes inasmuch as there were four persons receiving a score of 25 and four persons receiving a score of 26. In a case like that the mode may be said to be $25\frac{1}{2}$.

In many respects it is better to compute the mode from the frequency distribution rather than from an order distribution. When we start with a frequency distribution, the mode is taken as the middle of the step interval in which the largest number of scores fall. Thus, in the intelligence test there were 12 scores between 125 and 128, so the mode might well be said to be $126\frac{1}{2}$. In the case of the ethical scores, a frequency distribution is based on a step interval of 1 and so shows just what the order distribution showed. In the case of the religious scores, the mode would fall at $74\frac{1}{2}$, there being 6 scores between and including 70 and 79.

Interpretation: The mode is a rough approximation to the center of any group. It is the most common score. It shows the high point of the curve in a graphic distribution. The mode is not very frequently used in statistical computations. Occasionally it is the most important measure. This was true in the case of one group who were asked to rate on a scale a series of questions which they might be interested in discussing. The fact that the largest number of persons rated a certain question ten, and another question 5 was significant in guiding the discussion. It was probably more important to select the most frequent rating as indicating the value of that question for discussion than to select the average ratings which would have been

affected by some few people who may have liked or disliked the question to an unusual extent.

4. Other Significant Points

It is frequently desirable to summarize the location of other points within a group than simply those falling at or near the middle. As well as finding the midpoint of a distribution, it is interesting to note the quarter points, the tenths, and sometimes even the hundredths of the total distribution.

a. Q_1, *First Quartile, Lower Quartile*

Purpose: Q_1 is useful in showing the point below which one-quarter of the measures fall, and above which three-quarters of the measures may be found. A line drawn through Q_1 would separate the lowest 25 per cent of the group from the rest. This is frequently desirable when the purpose is to divide a group into four sections or to discover the extreme cases.

Computation: Q_1 may be found in two ways just as the median is found. It may be approximated by starting with an order distribution and counting up from the bottom one quarter of the total number of cases. Given 25 cases as in the case of the ethical scores, for example, this would mean 25 divided by 4 or $6\frac{1}{4}$ cases. Counting up $6\frac{1}{4}$ on the ethical scores we find that the sixth score is 20, and the seventh score is 20, so Q_1 will be 20. In the case of the religious scores, counting up $6\frac{1}{4}$ we come to a point quarter way between 41 and 45 which would be about 42. In the case of the intelligence scores Q_1 would fall between 125 and 126, but nearer 125.

Starting with a frequency distribution, the same procedure would be followed which was followed in the case of the median except that instead of counting up to half of the total number of scores, for Q_1, it would be necessary only to count up to one-fourth of the total number of scores. Where there are 25 cases, this means that we will count until we have used up $6\frac{1}{4}$ scores. Then $Q_1 = B + \frac{r \times s}{f}$, where B stands for the beginning of the step interval within which Q_1 falls, r stands for the remainder yet to be included to make a complete quarter of the number of cases, f stands for the frequency corresponding to the interval in which Q_1 falls, and s is the size of the interval. For illustration see Table IV.

Between Q_1 and the lowest score would be found roughly one-quarter of the cases if many had been measured. Between Q_1 and the median lie the second quarter. Any score below Q_1 is in the lowest 25 per cent. The distance between the lowest score and Q_1 in terms of actual score is usually much greater than the distance between Q_1 and the median.

b. Q_3, Third Quartile, Upper Quartile

Purpose: To find a point above which one-quarter of the cases will fall and below which three-quarters may be found.

Computation: Exactly the same as Q_1 except that instead of counting up one-fourth of the total number of cases, three-quarters are included (see Table IV).

Interpretation: Q_3 corresponds exactly to Q_1 except that it represents the other end of the scale. Sometimes the terminology is changed and Q_1 represents the high point, and Q_3 the low point. In any case, the middle half of all the scores lies between these two measures. A score above the upper quartile is in the best 25 per cent. One below the lower quartile is in the poorest 25 per cent.

c. Deciles

Purpose: In the same manner in which the quartile points Q_1, M, and Q_3 have been located, it is possible to locate the points which would divide up the total distribution into tenths.

Computation: Given 25 cases, it would be necessary to count up $2\frac{1}{2}$ cases to find the first decile point, 5 cases to the second, $7\frac{1}{2}$ to the third, etc.

Interpretation: Decile points are particularly useful in establishing norms. An individual who finds that his score fell between the fifth and the sixth decile knows that there were at least 60 per cent of the group who made scores poorer than his and at least 30 per cent who made scores better than his. Again it should be emphasized that the distance in actual score between the decile points at the extremes is far greater than the distance in actual score between the decile points near the center of the scale.

d. Percentiles

Purpose: Like the quartiles and the deciles, percentiles mark off a total distribution into points below which a certain per cent of the group may be found. Percentiles are useful in many cases for comparing scores. A score of 60 in a test of trustworthiness, and a score of 500 in a test on Bible knowledge are not comparable. No one knows what these mean, but if the first is stated as a percentile score of 30, meaning that 30 per cent of the

pupils made a lower score than 60 in trustworthiness, and the Bible score is stated as a percentile score of 22, meaning that only 22 per cent of the same pupils made a score less than 500, then it is fair to say that in comparison with his group, the pupil was better in his trustworthiness than in his Bible knowledge. Percentile scores are not quite so accurate for such comparisons as are the scores in terms of S.D. units described on pp. 152–153 because the distance in score between a percentile score of 5 and one of 10 is far greater than the distance between one of 55 and one of 60. In any normal distribution people are found bunched together near the middle, and to surpass 10 per cent of them there is easier than to surpass 1 per cent at the high or the low end.

Computation: Percentiles are ordinarily computed only when a large number of scores have been secured. If based on 1,000 scores, then the point below which 10 scores fall would be the one percentile point; the point below which 500 scores fall would be the 50 percentile point or median, and 25 and 75 percentiles are Q_1 and Q_3 respectively.

Interpretation: Any percentile score may be interpreted as meaning the per cent of scores, among all those upon which the distribution was based, falling below the given score. Zero would be the lowest, and 50 the average. The distance between percentile scores of 40 and 60, so far as actual score difference is concerned, is usually very slight.

e. Marks

Purpose: It has long been traditional in school examinations to give marks such as: excellent, good, fair, poor, or *A*, *B*, *C*, or I, II, III. These marks mean a wide variety of things in different situations. Sometimes they express simply the teacher's approval or disapproval based on the thousand subtle causes which might build up such an attitude. Sometimes they mean conformity to an arbitrary standard of achievement. Present tendencies in education are toward increasing the meaningfulness of such marks by interpreting them in terms of the total distribution of scores.

Computation: If the group is large and fairly normal with respect to the ability being marked, it may safely be assumed that there are a few at each extreme who deserve very high or very low marks, and that the

great majority belong in the middle of the ability score. The following example represents a customary practice when there are five marks to be assigned. If such a marking system is used with many pupils, it is on the whole true that the distance in ability between the average *A* and the average *B* is equal to the distance between the average *B* and the average *C*, which again in turn is equal to the distance between a *C* and a *D*, or a *D* and an *F*.

Suppose *A* means within the best 7 per cent
B means within the next 23 per cent
C means within the middle 40 per cent
D means within the next 23 per cent
F means within the lowest 7 per cent.

Then with 25 cases there would be approximately 2 *A*'s, 6 *B*'s, 9 *C*'s, 6 *D*'s, and 2 *F*'s. This would put the passing mark for the intelligence test in Table IV at 120, for the ethical test at 17, and for the religious test at 30. Of course the number of cases in the illustrations is far too small to justify rigid adherence to such a marking scale.

Modification would be especially desirable if the groups were not normal. If only boys and girls from better homes, or with better education were selected, there might well be more *A*'s and *B*'s, with fewer *D*'s and *F*'s. If the better boys and girls were excused from the test, or had been selected for a special class, that would lower the number of high marks, and increase the number of *D*'s and *F*'s.

Interpretation: Marking systems based on the distribution curve are at least clear in meaning. An *A* means within the best 3, 4, 7, or 10 per cent, depending on the scale used; a *C* distinctly means within the great middle section. This is not what marks mean at present when given by some teachers. There is room for a great deal of argument as to what marks ought to mean. The principal point of emphasis is that marks do not mean anything clearly and it is better to use marks which have an evident and unmistakable meaning even though that be not the very best conceivable meaning, than to continue the policy of having marks mean everything and nothing.

It may be pointed out that marks, upon this basis or any other, are no more reliable than the measuring instruments upon which they are based. If the test is no more reliable than the old essay examinations, or than most half-hour examinations of any sort, then an *A* might, if the test were repeated, become a *B*, a *C*, or even an *F*.

5. Measures of Variability. *a. Range*

Purpose: To give a general idea of variation.

Computation: Subtract the lowest score from the highest.

Interpretation: The statement of range, especially when the lowest and highest scores are mentioned, helps to interpret any individual score. A score of 70 on a test of caution means comparatively little unless it be stated that the total range of 1,000 scores was from 20 to 85. It is usually wise to state the number of cases in connection with the range.

b. The Average Deviation

Purpose: If it is known that class *A* made an average score of 127 on an intelligence test and that class *B* also made an average score of 127, it is not at all certain that the two groups are alike. It may be, as pointed out on page 184, that almost everyone in group *A* made a score not lower than 126 or higher than 128 whereas in group *B* the scores may have ranged from 50 to 200. If both were to be graphed on a single scale, group *A* would show a very high, narrow "skyscraper" distribution, whereas group *B* would be a low widespread "tabernacle" type of distribution. This difference may be expressed statistically in terms of a measure of the spread of the distribution away from the mean, called the average deviation.

Computation: To find the average deviation it is necessary to know the mean. Next, find the distance between each score and the mean, then average these deviations. Thus, in Table IX Ben scored 127 on the intelligence test which is 1.6 units away from the average of 128.6. The deviation for Clarence is 2.4, and so on down the line. When these deviations are added they yield a total of 141.2 or an average of 5.6. The average deviation for the ethical scores was 3.1 and that for the religious scores 18.9, indicating that there was very much greater variability among the scores on the religious test. On the ethical test the scores were very close together.

Interpretation: Average deviations between groups can be compared only when groups have taken the same test. It would not be at all true that this group varies six times as much in its religious knowledge as in its ethical ability. It is true that the

TABLE IX. — AVERAGE DEVIATION AND STANDARD DEVIATION

	Intelligence			Ethical score			Religious score		
Pupil	Score	Deviation	Deviation^2	Score	Deviation	Deviation^2	Score	Deviation	Deviation^2
1. Ben.....	127	1.6	2.56	25	2	4	58	5	25
2. Clarence	131	2.4	5.76	27	4	16	67	4	16
3. Dan....	127	1.6	2.56	26	3	9	71	8	64
4. Donald..	126	2.6	6.76	25	2	4	40	23	529
5. Dorothy	125	3.6	12.96	18	5	25	75	12	144
6. Florence	126	2.6	6.76	20	3	9	81	18	324
7. Fred....	140	11.4	120.96	23	0	0	94	31	961
8. Gordon..	127	1.6	2.56	24	1	1	86	23	529
9. Harold..	140	11.4	129.96	26	3	9	53	10	100
10. Helen...	128	0.6	0.36	21	2	4	41	22	484
11. Howard.	124	4.6	21.16	20	3	9	73	10	100
12. James...	132	3.4	11.56	28	5	25	50	13	169
13. Jeanne..	128	0.6	0.36	25	2	4	45	18	324
14. Jessie...	126	2.6	6.76	26	3	9	37	26	676
15. John....	146	17.4	302.76	27	4	16	74	11	121
16. Margaret	131	2.4	5.76	25	2	4	95	32	1,024
17. Marion..	129	0.4	0.16	22	1	1	90	27	729
18. Marvin.	128	0.6	0.36	24	1	1	62	1	1
19. Oscar...	117	11.6	134.56	16	7	49	32	31	961
20. Samuel..	126	2.6	6.76	15	8	64	54	9	81
21. Sarah...	122	6.6	43.56	17	6	36	29	34	1,156
22. Selma...	142	13.4	169.56	26	3	9	98	35	1,225
23. Theron..	137	8.4	70.56	27	4	16	77	14	196
24. Tony....	105	23.6	556.96	23	0	0	19	44	1,936
25. Walter..	125	3.6	12.96	19	4	16	74	11	121
Totals....	3,215	141.2	1,631.04	575	78	340	1,575	472	11,996
Means....	128.6	5.6	65.24	23	3.1	13.6	63	18.9	479.84
Square root	....	....	8.1	..	..	3.7	..	..	21.9
	A. D. = 5.6 S. D. = 8.1			A. D. = 3.1 S. D. = 3.7			A. D. = 18.9 S. D. = 21.9		

scores vary, by an amount six times as great, but that may be due purely to differences in the units in which the two tests are scored. For comparison, units must be alike. For example, if it is known that within the classes of one church school the average age deviation is 2 months, whereas in another school it is 2 years, one can be reasonably certain that it is easier to teach in the first school. At least so far as age is concerned the group would be more homogeneous in the school where the average deviation was small.

c. Standard Deviation

Purpose: The standard deviation is much more useful statistically than is the average deviation. Like the average deviation, it is a measure of spread and variation. Like the average deviation it gives a statistical measure which is based on all the measures and which is easily calculated. It has certain advantages that the average deviation does not have. It has a meaning in terms of the normal curve, that is, it measures the distance between the mean and the point at which the curve begins to turn from a convex top to a concave top. Thus, in Table VIII the distance marked σ corresponds to the standard deviation of that distribution. The standard deviation is useful in computing other measures, especially correlations, and it can be shown that this measure fluctuates least with variations in sampling.

Computation: Turning to Table IX, the method for computing the standard deviation is seen to be based on the same process as that which was used for the average deviation. The only difference is that in the case of the standard deviation each score deviation must be squared. Then the column is added, the average found, and the square root of the average gives the standard deviation. The standard deviation is sometimes called the mean square deviation. It is usually larger than the average deviation. In cases where an easy approximation to the standard deviation is desired, it is possible to use the nearest whole number if the mean represents a fraction. Thus, in the case of the intelligence scores, it might be possible to use a mean of 129 in which case all deviations would have been integers to be squared and added. This computation, like many others suggested in this chapter, will be facilitated by the use of squares and square roots from Table XXVIII in the Appendix.

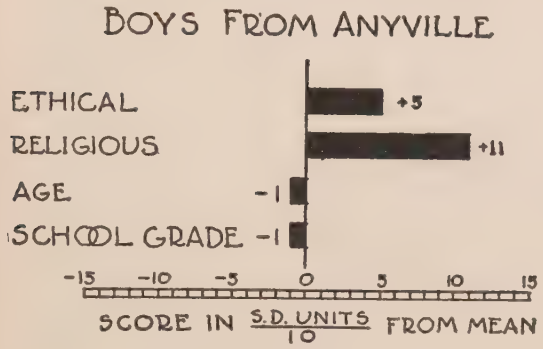
Interpretation: The standard deviation has several names. It has already been suggested that it is sometimes called the mean square deviation. It is also called sigma, and it may be indicated by a σ. In the formulae presented in this chapter it is symbolized by S.D. It is the standard measure of variation within the group. It is a distance along the base line of the graphic distribution. Usually if the data are approximately normal, and one standard deviation be laid off on either side of the mean, it will be found, as shown in Table VIII, that the area of the curve thus included represents about 68 per cent or roughly two-thirds of the total number of measures. Thus knowing that the mean for the religious scores is 63 and the standard deviation 22, it is possible to say with reasonable accuracy that two-thirds of the scores lie between 63 − 22 or 41 at the lower interval, and 63 + 22 or 85 at the high level. In the case of the ethical scores a small standard deviation of about 4 and the mean of 23 indicates that two-thirds or more of the cases could be found between 19 and 27.

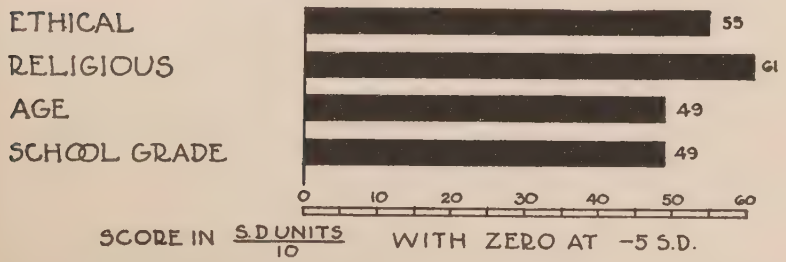
The standard deviation serves many other purposes besides simply measuring variability.[1] One of its most important functions is to serve as a unit of measure which will be comparable from one set of scores to another in the same group. In one research undertaking, boys from 50 cities had been given religious and ethical tests. It was known that the boys from Anyville made a score of 14 on the religious test and 36 on the ethical test. Were these boys better at religious ideas or in ethical judgment? Were they as good as boys their age ought to have been? From such data it would be impossible to say. The difference might be one of the scoring units employed by the different tests. There is no probability at all that a score of 20 on one test means the same thing as a score of 20 on another test. Suppose, however, that the score for each boy be stated in terms of its distance from the mean or from an assumed zero point in terms of standard deviation units. Suppose we know that the

[1] For a use of S. D. to measure variability leading to significant conclusions about the influences forming character, see Hartshorne and May, "Testing the Knowledge of Right and Wrong," Sixth Article, *Religious Education*, May, 1927.

average Anyville score when compared with the distribution of several thousand boys fell five-tenths of a standard deviation above the mean in the case of the ethical test and fell one and one-tenth S.D.'s above the mean in the case of the religious test, whereas in age and school grade the boys were one-tenth of an S.D. below the mean, that situation could be shown graphically

Table X. — Use of Standard Deviation in Comparing Scores





as in Table X. Turning the scores into S.D. units made it possible to compare the score which a boy made on one trait with scores based on a very different trait, and to show where they stood with reference to the country as a whole. This may be made even clearer by measuring not from the middle of the distribution curve, but from the low extreme. As stated above, the curve really runs out to infinity, but let us take a zero point, five S. D.'s below the mean. The relative scores are shown in the

second part of Table X. This graph gives a fairer impression of the relative values than does the distance measured from the mean when no zero point is indicated.

If it is desired to state these scores numerically, the unit chosen is one-tenth of a standard deviation. This saves the use of decimal points and gives us a scale in which since the zero is at − 5 S.D. 100 is practically perfect, and 50 is average. This is the plan employed in McCall's T scale.[1]

Standard deviations may also be used as a measure of reliability. This is discussed later on pages 237–243.

d. Q

Purpose: Q is only a nickname for a measure properly christened semi-interquartile range. Like the average deviation and the standard deviation it is a measure of spread. It belongs in the family with the median also, because it does not take account of extreme measures but counts to certain positions rather than averaging all scores. Many people prefer to use Q rather than standard deviation when expressing the amount of range within a group because Q is so much easier to compute.

Computation: $Q = \frac{Q_3 - Q_1}{2}$. The steps in finding it are: first, count up to Q_1, the lower quartile. Then count up to Q_3, the point below which three-fourths of the measure fall, then subtract Q_1 from Q_3 and divide the result by 2. Table IV indicates how this was done in connection with frequency distributions for the intelligence, ethical, and religious scores It can be done using Q_3 and Q_1 as approximated from an order distribution also.

[1] See McCall, "How to Measure in Education," pp. 272–307.

The procedure in many books on statistics is to express distance from the mean, rather than from a zero point 5 S. D.'s below. To use such tables, find for each score the per cent of the total number of scores lying between the given score and the median. Transmute the per cent (which represents a portion of the area of the curve) into S. D. distance from the mean. Multiply it by ten to obtain S. D. units as used here. If the given score is below the median, subtract the result from 50. If it is above the median add the result to 50.

Table XXXIII in the Appendix gives a more direct method. Find first the per cent of all pupils making a higher score than the one in question. Then read directly from the table the corresponding score in S. D. units. For transmuting a large series of scores, make a table giving for each raw score its equivalent in S. D. units.

Interpretation: Q represents the distance which, if laid off on both sides of the median, will include half of the cases. Thus, in the case of the intelligence scores we know that the median is approximately 128. Q is 3. We can, therefore, say that half of the cases fall between a score of 125 and a score of 131. Where Q is large, the spread is wide and the groupings are not very homogeneous. When Q is small, the group is compact and alike in the trait being measured. In the normal distribution Q is equal to the average deviation, and to the P.E. or probable error.[1] Sometimes Q, P.E., and A.D.[2] are used instead of S.D. as a unit along the base line of the curve, by which to compare scores from different tests.

6. Measures of Relationship. *a. Graphic Measures*

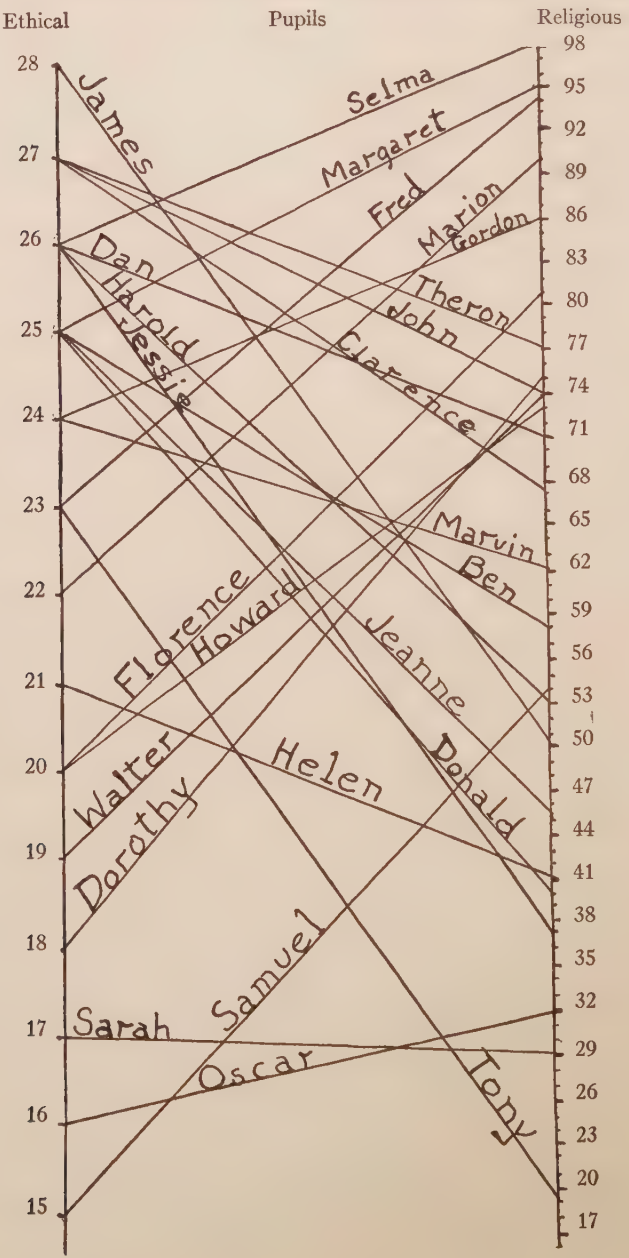
Purpose: Graphic measures of relationship are sometimes used to show how the order of score in one trait compares with the order of score in the second trait. In the accompanying table (Table XI), lines have been drawn to connect each pupil's religious score with his ethical score. It is seen that some who stood near the top in one distribution fell near the bottom in the other. If there were a perfect positive relationship between the scores in the two traits every line would run right straight across, more or less as do now the lines for Sarah, Helen, and Marvin. If all the scores ran as do those for James, Jessie, Tony, Samuel, Dorothy, and Florence, a negative relationship would be seen to exist. That is, persons high on one test would be just as low on the second as they were high in the first. When the lines go every which way, as they do in Table XI, the relationship is very nearly zero, or chance.

b. Correlation

Purpose: The most common and perhaps most overworked measure of relationship is correlation. It aims to tell how closely one series of scores or measures corresponds to another series of scores and measures based on the same pupils or situations. It is the aim of correlation to answer such questions as the following: "How closely is religious knowledge related to intelligence?" "What is the relationship between Sunday-school attendance and

[1] See p. 237. [2] See p. 191.

TABLE XI. — GRAPH OF RELATION BETWEEN ETHICAL AND RELIGIOUS SCORES



juvenile delinquency?" "Do people who are very much interested in art and music make good thinkers?" "Is there a tendency for people to become more conservative as they grow older?" The correlation coefficient between the first of each of these pairs of measures and the second would be of considerable aid in answering the questions.

Computation: Only two methods for computing correlation will here be suggested.[1] The first is based on the rank of pupils. The steps in its computation are as follows:

1. Make a rank distribution for each trait.
2. Find the difference between a pupil's rank in one measure, and his rank in the other.
3. Square each difference.
4. Add these squares.
5. Fill in the following formula:

Rho (ρ) equals the correlation coefficient as obtained by this rank method. ΣD^2 equals the sum of the deviations squared as found in steps one to three. n equals the number of cases. Then

$$\text{Rho} = \frac{1 - (6 \times \Sigma D^2)}{n(n^2 - 1)}$$

This is illustrated in Tables XII, XIII, and XIV.

Rho, however, is not exactly equivalent in its assumptions to r the usual Pearson, or "product moment" coefficient of correlation. To correct for this slight error, Table XXIX has been prepared. The value of r may be read corresponding to any computed value of rho.

The rank method is less accurate and less desirable than one of the standard methods for computing r directly. r is sometimes called Pearson's coefficient of correlation after the great biometrician who developed it. To compute r directly the following steps may be taken:

1. Square each score in each trait.
2. Find the product of every pupil's score in the first trait, and his score in the second.
3. Add the five columns resulting, that is, pupils' scores in measure A, their scores in measure B, the square of the A scores, the square of the B scores, and the products or AB scores.
4. Find the average for each column. This must be carried out very exactly.

[1] Both of these methods assume a linear regression. Sometimes this yields too low a result if the real relationship is curvilinear. For the computation or interpretation of eta (η) see GARRETT, "Statistics in Psychology and Education," pp. 203–211, Longmans, Green, and Co., New York, 1926.

TABLE XII. — RANK METHOD CORRELATION BETWEEN INTELLIGENCE AND ETHICAL SCORE

Pupil	Rank in intelligence	Rank in ethical score	Difference	Square of difference
1. Ben	14	10½	3½	12.25
2. Clarence	7½	3	4½	20.25
3. Dan	14	6½	7½	56.25
4. Donald	17½	10½	7	49
5. Dorothy	20½	22	1½	2.25
6. Florence	17½	19½	2	4
7. Fred	3½	15½	12	144
8. Gordon	14	13½	½	0.25
9. Harold	3½	6½	3	9
10. Helen	11	18	7	49
11. Howard	22	19½	2½	6.25
12. James	6	1	5	25
13. Jeanne	11	10½	½	0.25
14. Jessie	17½	6½	11	121
15. John	1	3	2	4
16. Margaret	7½	10½	3	9
17. Marion	9	17	8	64
18. Marvin	11	13½	2½	6.25
19. Oscar	24	24	0	0
20. Samuel	17½	25	7½	56.25
21. Sarah	23	23	0	0
22. Selma	2	6½	4½	20.25
23. Theron	5	3	2	4
24. Tony	25	15½	9½	90.25
25. Walter	20½	21	½	0.25
Total	....	....	....	753.00

$$\text{Rho} = 1 - \frac{6 \times 753}{25(25^2 - 1)}$$

$$= 1 - \frac{4{,}518}{15{,}600}$$

$$= 1 - 0.29$$

$$= 0.71$$

By Table XXIX

$$r = 0.73$$

NOTE: It may be of assistance in calculating the column of squares to note that the square of any number (a) plus one-half, is equal to the product of that number and the next integer, plus one-quarter.

Thus

$$(a + \tfrac{1}{2})^2 = (a.[a + 1]) + \tfrac{1}{4}$$

The square of $4\frac{1}{2}$, for example, is $4 \times 5 + \frac{1}{4}$, or 20.25

TABLE XIII. — RANK METHOD CORRELATION BETWEEN INTELLIGENCE AND RELIGIOUS SCORE

Pupil	Rank in intelligence	Rank in religious score	Difference	Square of difference
1. Ben	14	15	1	1
2. Clarence	7½	13	5½	30.25
3. Dan	14	12	2	4
4. Donald	17½	21	3½	12.25
5. Dorothy	20½	22	1½	2.25
6. Florence	17½	6	11½	132.25
7. Fred	3½	3	½	0.25
8. Gordon	14	5	9	81
9. Harold	3½	17	13½	182.25
10. Helen	11	20	9	81
11. Howard	22	11	11	121
12. James	6	18	12	144
13. Jeanne	11	19	8	64
14. Jessie	17½	22	4½	20.25
15. John	1	9½	8½	72.25
16. Margaret	7½	2	5½	30.25
17. Marion	9	4	5	25
18. Marvin	11	14	3	9
19. Oscar	24	23	1	1
20. Samuel	17½	16	1½	2.25
21. Sarah	23	24	1	1
22. Selma	2	1	1	1
23. Theron	5	7	2	4
24. Tony	25	25	0	0
25. Walter	20½	9½	11	121
Total	....	....	....	1,142.5

$$\text{Rho} = 1 - \frac{6 \times 1{,}142.5}{25(25^2 - 1)}$$

$$= 1 - \frac{6{,}855}{15{,}600}$$

$$= 1 - 0.44$$

$$= 0.56$$

From Table XXIX

$$r = 0.58$$

TABLE XIV. — RANK METHOD CORRELATION BETWEEN ETHICAL SCORE AND RELIGIOUS SCORE

Pupil	Rank in ethical score	Rank in religious score	Difference	Square of difference
1. Ben	10½	15	4½	20.25
2. Clarence	3	13	10	100
3. Dan	6½	12	5½	30.25
4. Donald	10½	21	10½	110.25
5. Dorothy	22	22	0	0
6. Florence	19½	6	13½	182.25
7. Fred	15½	3	12½	156.25
8. Gordon	13½	5	8½	72.25
9. Harold	6½	17	10½	110.25
10. Helen	18	2	16	256
11. Howard	19½	11	8½	72.25
12. James	1	18	17	289
13. Jeanne	10½	19	8½	72.25
14. Jessie	6½	22	15½	245.25
15. John	3	9½	6½	42.25
16. Margaret	10½	2	8½	72.25
17. Marion	17	4	13	169
18. Marvin	13½	14	½	0.25
19. Oscar	24	23	1	1
20. Samuel	25	16	9	81
21. Sarah	23	24	1	1
22. Selma	6½	1	5½	30.25
23. Theron	3	7	4	16
24. Tony	15½	25	9½	90.25
25. Walter	21	9½	11½	132.25
Total	....	....	....	2,352

$$\text{Rho} = 1 - \frac{6 \times 2{,}352}{25(25^2 - 1)}$$

$$= 1 - \frac{14{,}112}{15{,}600}$$

$$= 1 - 0.90$$

$$= 0.10$$

From Table XXIX

$$r = 0.11$$

5. Substitute in the following formula:

$$r = \frac{Mn_{AB} - Mn_A \times Mn_B}{\sqrt{(Mn_{A2} - [Mn_A]^2)\,(Mn_{B2} - [Mn_B]^2)}}$$

where

Mn_{AB} means the average of the columns of products, the AB columns
Mn_A means the average of the column of scores in trait A
Mn_B means the average of the column of scores in trait B
Mn_{A2} means the average of the column of squares of scores in trait A
Mn_{B2} means the average of the column of squares of scores in trait B
$[Mn_A]^2$ means the square of Mn_A
$[Mn_B]^2$ means the square of Mn_B

This can best be understood by following through the computations in Tables XV, XVI, and XVII.[1]

Other methods for computing r directly may be found in other texts on statistics. This particular method is chosen because if the number of cases is small, it can be very directly handled without the bother of making out a scatter diagram, and because if the number of cases is large, it can be handled on an adding machine with far greater ease than it can be tabulated. When intercorrelations are needed, as is shown in Table XXII, very little additional work has to be done.

One modification of the above procedure may do a great deal to save work. Correlation takes account only of spread within the group and not of the actual size of the scores. It is, therefore, possible to subtract from any series of scores a figure such that the lowest score in the group becomes one. This will not affect the correlation at all, and will frequently greatly reduce the amount of labor involved in finding squares or products. If the numbers are still large for convenient handling, they may be divided by 2, 3, 5, 10, or any number to bring them into convenient range. A fraction should be dropped, only the nearest integer being used. Persons familiar with other methods of finding correlation will recognize that this is equivalent to grouping the scores in frequency distribution intervals of 2, 3, 5, 10, or any given number. It is good practice not to use less than ten such intervals, or stated in terms of the formula here presented, not to divide

[1] NOTE: For a study of the theory of correlation and its mathematical meaning, running beyond the suggestions on interpretation given later in this chapter, see RUGG, "Statistical Methods Applied to Education," or discussions in any of the statistical texts listed on p. 172. Each has one or more chapters devoted to correlation. For a comparison of this particular formula with the others in common use, see SYMONDS, "Variations of the Product-Moment (Pearson) Coefficient of Correlation," *Jour. Ed. Psy.*, October, 1926, Vol. XVII, p. 458. If it is desired to study relationships between traits measured in uncertain or irregular units, the coefficient of contingency "C" is preferable to "r." See discussions in Kelly, Yule, or Rietz.

TABLE XV. — CORRELATION BETWEEN INTELLIGENCE AND ETHICAL SCORES[1]

Pupil	Intelligence scores			Ethical scores			Product of reduced scores
	Original	Reduced	Square	Original	Reduced	Square	
1. Ben.....	127	14	196	25	11	121	154
2. Clarence..	131	16	256	27	13	169	208
3. Dan.....	127	14	196	26	12	144	168
4. Donald...	126	13	169	25	11	121	143
5. Dorothy..	125	13	169	18	4	16	52
6. Florence..	126	13	169	20	6	36	78
7. Fred.....	140	20	400	23	9	81	180
8. Gordon...	127	14	196	24	10	100	140
9. Harold...	140	20	400	26	12	144	240
10. Helen....	128	14	196	21	7	49	98
11. Howard...	124	12	144	20	6	36	72
12. James....	132	16	256	28	14	196	224
13. Jeanne...	128	14	196	25	11	121	154
14. Jessie.....	126	13	169	26	12	144	156
15. John.....	146	23	529	27	13	169	299
16. Margaret.	131	16	256	25	11	121	176
17. Marion...	129	15	225	22	8	64	120
18. Marvin...	128	14	196	24	10	100	140
19. Oscar.....	117	9	81	16	2	4	18
20. Samuel...	126	13	169	15	1	1	13
21. Sarah....	122	11	121	17	3	9	33
22. Selma....	142	21	441	26	12	144	252
23. Theron...	137	19	361	27	13	169	247
24. Tony.....	105	3	9	23	9	81	27
25. Walter...	125	13	169	19	5	25	65
Total.....	...	363	5,669	..	225	2,365	3,457
Means.....	...	14.52	226.76	..	9.0	94.6	138.28

$$r = \frac{138.28 - (14.52 \times 9.0)}{\sqrt{(226.76 - 14.52^2)(94.6 - 9.0^2)}}$$

$$= \frac{138.28 - 128.68}{\sqrt{(226.76 - 210.83)(94.6 - 81.0)}}$$

$$= \frac{9.60}{\sqrt{15.93 \times 13.6}}$$

$$= \frac{14.65}{9.60}$$

$$= 0.66$$

[1] These intelligence scores are reduced, to make the computation easier, by subtracting 100, and dividing by two, keeping no fractions.

The ethical scores have been reduced, for ease of computation, by subtracting 14 from each.

TABLE XVI. — CORRELATION BETWEEN INTELLIGENCE AND RELIGIOUS SCORES[1]

Pupil	Intelligence scores			Religious scores			Product of reduced scores
	Original	Reduced	Square	Original	Reduced	Square	
1. Ben.....	127	14	196	58	9	81	126
2. Clarence..	131	16	256	67	10	100	160
3. Dan.....	127	14	196	71	11	121	154
4. Donald...	126	13	169	40	5	25	65
5. Dorothy..	125	13	169	75	12	144	156
6. Florence..	126	13	169	81	13	169	169
7. Fred.....	140	20	400	94	16	256	320
8. Gordon...	127	14	196	86	14	196	196
9. Harold...	140	20	400	53	8	64	160
10. Helen....	128	14	196	41	5	25	70
11. Howard...	124	12	144	73	12	144	144
12. James....	132	16	256	50	7	49	112
13. Jeanne...	128	14	196	45	6	36	84
14. Jessie.....	126	13	169	37	4	16	52
15. John.....	146	23	529	74	12	144	276
16. Margaret	131	16	256	95	16	256	256
17. Marion...	129	15	225	90	15	225	225
18. Marvin...	128	14	196	62	9	81	126
19. Oscar.....	117	9	81	32	3	9	27
20. Samuel...	126	13	169	54	8	64	104
21. Sarah....	122	11	121	29	7	49	77
22. Selma....	142	21	541	98	17	289	357
23. Theron...	137	19	361	77	12	144	228
24. Tony.....	105	3	9	19	1	1	3
25. Walter...	125	13	169	74	12	144	156
Total.....	...	363	5,669	..	244	2,832	3,803
Means.....	...	14.52	226.76	..	9.76	113.28	152.12

$$r = \frac{152.12 - (14.52 \times 9.76)}{\sqrt{(226.76 - 14.52^2)(113.28 - 9.76^2)}}$$

$$= \frac{152.12 - 141.72}{\sqrt{(226.76 - 210.83)(113.28 - 95.26)}}$$

$$= \frac{10.40}{\sqrt{15.93 \times 18.02}}$$

$$r = \frac{10.4}{17.0}$$

$$= 0.61$$

[1] Intelligence scores are here reduced by subtracting 100 and dividing by two, using the nearest integer.

Religious scores are here reduced by subtracting 15 and dividing by 5, using the nearest integer.

TABLE XVII. — CORRELATION BETWEEN ETHICAL SCORES AND RELIGIOUS SCORES

Pupil	Product of reduced scores
1. Ben	99
2. Clarence	130
3. Dan	132
4. Donald	55
5. Dorothy	48
6. Florence	78
7. Fred	144
8. Gordon	140
9. Harold	96
10. Helen	35
11. Howard	72
12. James	98
13. Jeanne	66
14. Jessie	48
15. John	156
16. Margaret	176
17. Marion	120
18. Marvin	90
19. Oscar	6
20. Samuel	8
21. Sarah	21
22. Selma	204
23. Theron	156
24. Tony	9
25. Walter	60
Total	2247
Mean	89.88

From Tables XXV and XXVI

Mean of ethical scores	9.0
Mean of squares	94.6
Mean of religious scores	9.76
Mean of squares	113.28

$(94.6 - 9.0^2) = 13.6$

$(113.28 - 9.76^2) = 18.02$

Hence

$$r = \frac{89.88 - (9.0 \times 9.76)}{\sqrt{13.6 \times 18.02}}$$

$$= \frac{2.04}{15.7}$$

$$= 0.13$$

the original scores by a number which will make the range less than 10. In the case of the intelligence scores presented in Table XV, the scores have been reduced by subtracting 100 and dividing by 2. From the ethical scores 14 was subtracted. The religious scores were reduced by subtracting 15, and dividing by 5. In many cases this makes the task of correlation finding almost a matter of simple multiplication table work.

Interpretation: Four methods will be suggested for the interpretation of correlation. The first is in terms of its usual meaning. A correlation of 1.00 means perfect agreement. It would be obtained, for example, between the height of a large number of people measured in inches and the same height measured with equal accuracy in terms of a metrical scale. Each person would stand in one list exactly where he stood in the other. A correlation of zero means pure chance, no relationship whatever. Such a correlation might be expected between the number of words in a person's vocabulary and the amount of pigment in his hair. Some blondes would talk more and some would talk less. The correlation would be, in all probability, very close to zero. It is probable that the correlation between intelligence and happiness in life is not far from zero. If the correlation is negative, it means that people who are high in one trait have some tendency to be low in the other. The correlation between the number of times each pupil is sent to the superintendent's office for misbehavior and the rating which each would receive for good conduct would probably be a -0.80 or -0.90. The correlation between the amount of money spent on church buildings in a series of communities and the proportion of unchurched population would probably be negative although perhaps not higher than -0.50. Rugg states that as a result of his experience he is inclined to call correlations of less than 0.20 negligible or indifferent, correlations of 0.20 to 0.40 as being present but low, 0.40 to 0.60 as being marked, and above 0.70 as high. Many others would use a somewhat higher standard, being unwilling to regard a correlation as high unless it exceeded 0.80 or 0.85.

Correlation may be interpreted in the second place in terms of the error of prediction. Suppose we are given the intelligence scores of a group of pupils and on the basis of this we try to

predict what their ethical insight would be as measured by an ethical test, what will be the error of prediction? How largely will the result be what it would have been had we made a blind guess? The "predictive index"[1] affords an answer to this question. The predictive index is $1 - \sqrt{(1 - r^2)}$. In the case of our problem we found correlation between ethical judgment and intelligence to be 0.66. The predictive index gives us:

$$1 - \sqrt{1 - (0.66^2)} = 1 - 0.73 = 0.27$$

That would indicate that we have gone approximately 27 per cent of the way from sheer chance to perfect prediction. If this predictive index were zero, it would be a blind guess. If it were 1.00 we would hit it absolutely right every time. As it is the correlation of 0.66 means an improvement of about 0.27 over unguided guesswork. Table XXX gives a statement of the predictive index for various coefficients of correlation. From that it will be observed that a correlation in order to be half way between sheer chance and perfect prediction must be 0.86, and to go three-fourths of the way it must be above 0.95. Correlations as low as 0.60 are 80 per cent pure chance in their power to predict.

A third method for interpreting correlation is to consider the average displacement which would be caused if we were to predict one trait from our knowledge of the other. Table XXXI shows the displacement in tenths marked off along the base line of a distribution curve (after Otis). In the case of a correlation of 0.66 we can estimate that about 30 per cent of the scores will fall in the same tenth in one test that they did in the other; about 75 per cent will fall in the same tenth or be displaced by not more than one-tenth in either direction; about 95 per cent will be found within two-tenths of the distribution above or below the original score, whereas the remaining 5 per cent will probably be displaced by more than one-fifth of the range of scores. It would be disastrous to deal with individuals on the assumption

[1] Developed by BAILOR, "Content and Form in Tests of Intelligence," pp. 25–29, Teachers College, Columbia University, New York, 1924. It is equivalent to 1 minus the "coefficient of alienation" which tells how far from perfection the prediction is.

that a test which measured a trait with a correlation of only 0.66 between the test and the true value of the trait was really fair to individuals; 70 per cent of them would be seriously displaced. Of course, it must be remembered that these figures are only averages based on a normal group, and that there will be variations from them in the case of any given series of scores, and errors in the case of scores which stood near the extremes will, in general, be greater than errors among scores near the center of the distribution. The table does, however, give a general notion of the accuracy of prediction.

Probably the most useful general interpretation of correlation is in terms of the percentage of causal factors which the two traits have in common. Table XXXII worked out by Nygaard, the derivation for which is published in the *Journal of Educational Psychology* for February, 1926, indicates for any coefficient of correlation the per cent of causal factors which the two measures correlated, share. Thus, given a correlation of 0.66, we find that 47 per cent of the factors which cause persons to rate high in one test were also present, causing them to rate high in the second, and the same percentage of factors causing low scores in one test tended to cause low scores in the other. It is assumed that these causal factors are completely independent. This does not say that one score causes the other. There may be a correlation of 0.13 between religious knowledge and ethical knowledge, but that does not indicate that religious training is responsible for ethical knowledge even to the extent of 12 per cent. It may quite as well be that ethical knowledge is responsible for religious insight or that both are due to intelligence or home background or other factors.

Correction, for Attenuation: Occasionally it becomes desirable to know what the correlation between two traits would have been had the tests been much more reliable and free from chance errors. This can be determined by employing the correction for attenuation.

Computation: Let r_{AB} represent the true correlation between tests A and B.

Let $r_{A_1A_2}$ represent the self-correlation of test A.

Let $r_{B_1B_2}$ represent the self-correlation of test B.
Let $r_{A_1B_2}$ represent the obtained correlation between A_1 and B_2.
Let $r_{A_2B_1}$ represent the obtained correlation between A_2 and B_1, or
Let $r(AB)$, represent any obtained correlation between A and B.
Then

$$r_{AB} = \frac{\sqrt{(r_{A_1B_2})(r_{A_2B_1})}}{\sqrt{(r_{A_1A_2})(r_{B_1B_2})}} \text{ or, usually } = \frac{r(AB),}{\sqrt{(r_{A_1A_2})(r_{B_1B_2})}}$$

In Table XVIII is shown the true relationship between religious ideas and ethical judgment which would have been found had the tests for religious ideas and tests for ethical judgment been perfectly reliable. It is assumed that the reliability of the ethical test was 0.70 and the reliability of the religious test on a group of equal variability, was 0.80. From this it appears that the correlation with completely reliable tests would have been nearer 0.17 than the obtained correlation of 0.13.

Attenuation is based upon the assumption that the errors in each test are purely chance and are quite unrelated to one another. That is seldom true. The correction for attenuation usually gives a result which is too high. More on the basis of common sense than on the basis of statistical theory McCall has suggested that the true result be stated as the average between the figure obtained in the original correlation and that which is found to be the "true" correlation, using this correction. The formula as stated has led to numbers of absurd conclusions. It not infrequently produces correlations larger than 1.00. It would lead one to believe that if there are low correlations between two very unreliable tests or ratings that the true correlation between the tests or traits must be almost perfect. This is often not the case. Probably the errors are less serious when the reliabilities of the tests are rather high in the beginning. Reliability should not be confused here with validity. Correction for attenuation tends to indicate what the answer would be if

TABLE XVIII. — EXAMPLE OF CORRECTION FOR ATTENUATION

Correlation between religious and ethical scores = 0.13
Reliability of religious test = 0.80
Reliability of ethical test = 0.70

Then

$$\text{True correlation} = \frac{r_{re}}{\sqrt{(r_{e_1e_2})(r_{r_1r_2})}} = \frac{0.13}{\sqrt{0.70 \times 0.80}} = 0.17$$

the tests were perfectly reliable. The tests may still, of course, be reliable measures of something very different from the trait the test makers wanted to measure. It is a "true correlation" only in the sense of being truly reliable.

c. Regression

Purpose: Given the correlation between two sets of measures, it is possible to predict within certain limits the most probable score in one series from knowledge of an individual's score in another series. If it is known that the correlation between type of home and the child's ethical vocabulary is 0.80, it is possible to predict with reasonable accuracy what the most probable type of home is, for a child who makes a score on ethical vocabulary of 67 or 89. The coefficient of correlation itself, although it is more prominent now than the regression equation, was really devised as a means to this type of prediction.

Computation: Suppose it is desired to predict an individual's score in A from knowledge of his standing in trait B. The equation then becomes:

$$A = r_{AB} \times \frac{\text{S.D.}_A}{\text{S.D.}_B}(B - Mn_B) + Mn_A$$

Table XIX shows the application of this formula to the prediction of the most probable ethical score given an intelligence score or religious score.

TABLE XIX. — PREDICTION OF THE MOST PROBABLE ETHICAL SCORE FROM A GIVEN INTELLIGENCE SCORE OR RELIGIOUS SCORE[1]

Given: Intelligence score of 131

Let e be the desired ethical score.

$$e = 0.66 \times \frac{3.7}{8.1}(131 - 128.6) \div 23.$$

$$e = (0.66 \times 0.46 \times 2.4) \div 23$$

$e = 23.33$, the most probable score on the ethical test.

Given: Religious score of 41

Let e be the desired ethical score

$$e = 0.11 \times \frac{3.7}{21.9} \times (41 - 63) \div 23$$

$$e = (0.11 \times .17 \times -22) \div 23$$

$e = -0.41 \div 23 = 22.59$, the most probable score on the ethical test.

Thus, when the intelligence score is 131, we need only to substitute in the formula the correlation of 0.66 which was obtained between intelligence and ethical score and the standard deviations shown in Table IX which

[1] See p. 240 for the reliability of these results.

were 3.7 for ethical score, and 8.1 for intelligence. Multiply this by the difference between the score of 131 and the intelligence mean, which was 128.6, and add to the result the mean of the ethical scores, which was 23. So doing it appears that the most probable score on an ethical test for an individual who had an intelligence score of 131 is 23.33. In the same fashion it is shown that the most probable score for an individual who had a score of 41 on the religious test is 22.59 when it comes to the ethical test.

Interpretation: Regression is a matter of probability. In any given case the limits of error may be fairly large (see p. 240). It is based on the assumptions that the scores follow a normal distribution, and that the regression lines are straight and units all alike. It is subject to many limitations in practical use, especially when based only on a simple correlation. For example, the correlation between income and number of automobiles owned is probably 0.50. On this basis one might predict that the writer owns 0.8 of an automobile whereas, as a matter of fact, he owns not even a collegiate Ford. The most probable score is really the average score of all of the people who made a score similar to the one given in the trait which is used as the basis of prediction. In the automobile case, it would be the average number of automobiles owned by persons with the same income as the writer's.

d. Critical Score

Purpose: In connection with tests for vocational guidance it is frequently relatively unimportant to know what the correlation is between success in a certain test and success in a certain line of work. It is rather desired to locate points above which a person is almost certain to be successful and below which a person is almost certain to fail.

Computation and Interpretation: For the ordinary situation such critical scores can best be determined by trial and error. The accuracy of prediction may be stated in terms of the per cent of success. Thus, it may be said that if a critical score based on the combination of intelligence test scores and high school marks be set at a certain point in a college, then only 1 per cent of the persons falling below that would have been able to succeed in college, whereas 88 per cent of those who fell above it made

successful college careers. Again another point may be located so high up the line that it may be said with some certainty that only 3 per cent of those making scores above this point will ever drop out of college because of unsatisfactory work. In administration problems it is frequently desirable to set a single point which separates high chance of successful work from low chance. There is probably a fair positive correlation between the circulation of air in a room, and the mental efficiency of children working in the building. Yet it is better for administration purposes to state a point below which the results are almost certain to be undesirable and above which they are almost certain to be satisfactory. Sometimes a test maker may wish to validate a test not so much in terms of correlation, as in terms of critical score. He may say, "I don't know how this test correlates in general with the criterion, but I do know that any person making a score under 10 is so ignorant as to be socially dangerous, and any person making a score over 60 is so well trained as to be almost invariably socially useful." The number of errors per hundred in the case of each prediction should be stated.

e. Partial Correlation

Purpose: Partial correlation is used to find out the true relationship between one series of measures and another series, as it would have been had certain other measures been constant instead of variable. An investigator might find that the correlation between Sunday-school attendance and appearance in juvenile court is -0.60. It would look as though Sunday school tended to prevent delinquency. The careful experimenter immediately recognizes other causes which may be involved. He asks whether all these pupils came from the same type of home. Finding that they did not, he suggests that the homes be scored on a home-rating scale, and that that factor be, as the saying is, "partialled out," by use of the partial correlation technique. It may then appear that the correlation between Sunday-school attendance and juvenile delinquency, when the home situation is assumed to be just alike for each person, is only -0.15. That would mean that Sunday-school attendance had relatively little to do with delinquency when

home factors were constant. Or it might be that the partial correlation would jump to -0.80 or -0.90 indicating that if persons from equal home environments were studied, the influence of Sunday school, partly covered up by home differences before, would stand out.

It has been well said that the development of partial correlation represents one of the great advances of human science. It is possible, by its use, to study almost pure causal relationships, even in a complicated social situation. It does require large numbers of cases, if the results are to be very reliable.

Computation: The typical formula for partial correlation, when it is desired to know the real correlation between a and b, with c constant or parceled out, is

$$r_{ab \cdot c} = \frac{r_{ab} - (r_{ac} r_{bc})}{\sqrt{(1 - r_{ac}^2)\,(1 - r_{bc}^2)}}$$

In this formula

$r_{ab \cdot c}$ means the correlation between a and b, when c is constant.
r_{ab} means the correlation between a and b
r_{bc} means the correlation between b and c
r_{ac} means the correlation between a and c

Suppose the problem is to find out what the correlation would be between religious score and ethical score, if all the pupils taking tests had been of the same intelligence. We found a correlation of 0.13 in Table XVII which indicated that there was some tendency for high scores in one to go with high scores in the other, but that this tendency was very slight. How will it be when the pupils are equated for intelligence? The result worked out in Table XX is -0.35, indicating a fair tendency for pupils who did

TABLE XX. — PARTIAL CORRELATION BETWEEN ETHICAL SCORE AND RELIGIOUS SCORE WHEN INTELLIGENCE IS HELD CONSTANT

Taken from previous tables:

Correlation between ethics and religion	0.13
Correlation between intelligence and religion	0.61
Correlation between intelligence and ethics	0.66

$$r_{re \cdot i} = \frac{0.13 - 0.61 \times 0.66)}{\sqrt{(1 - 0.61^2)(1 - 0.66^2)}}$$

$$= \frac{0.13 - 0.40}{\sqrt{(1 - 0.37)(1 - 0.44)}}$$

$$= \frac{-0.27}{\sqrt{0.59}}$$

$$= -0.35$$

well on the religious test to do poorly on the ethical test, when all are alike in intelligence.

When there are a number of variables to be held constant, the process seems complicated, but it is really only a repetition over and over again of this simple formula. The steps in the procedure are:

1. Find the intercorrelations of each measured factor with every other.
2. Find the first-order partials. That is, take every possible pair of factors, and in the manner illustrated above hold a certain factor, for example c, constant.
3. Find the second-order partials. Take the results from step (2) and treat them as though they were the original correlations, and from each of them, partial out a second factor, for example d.
4. Find the third-order partials. Take the results from step (3) and treat them as though they were the original correlations, and from each of them, partial out a third factor, for example, e.

And so on. One at a time, each of the variables may be removed by the same formula as that given above. When it is used the first time, it contains zero-order correlations, r_{ab}, r_{az}, etc. The second time, in exactly the same places we may have $r_{ab \cdot c}$ and $r_{az \cdot c}$. The third time, after c and d are both removed, we may have $r_{ab \cdot cd}$ and $r_{az \cdot cd}$ but however far we go, the form for arithmetical work stays just the same.

The end result $r_{ab \cdot cdyz \cdot \ldots n}$ is the result of taking out first c, then d, then y, then z, then the next factor, and so on until n have been removed.

This procedure is again illustrated a little later [1] in connection with an administrative problem involving five variables, three of which are held constant while the relation is found between the other two.

Interpretation: The interpretation of partial correlation rests upon that of simple correlation. It shows how well one series of scores corresponds with another. This may be thought of in terms of displacements when prediction is made, or in terms of improvement over sheer chance, or in terms of per cent of common causes. The difference lies in the fact that other factors are held constant. The setting in which the relationship is studied is much more ideal for discovering truth. It is hard to see the effect of Sunday school on delinquency when such variables as different homes, different intelligence, different gangs, different public schools, different ancestry, all mix up together. But if one by one these are all separated out we come closer and closer to pure cause.

[1] See p. 224.

Partial correlation is subject to extraordinary difficulty in interpretation by those unacquainted with the assumptions involved in its development. A few comments may help. First, observe that although two factors may be shown to be related, causal connections are likely still to be in doubt. Did the first cause the second, or the second cause the first, or are both influenced by something else quite different? Next notice that when certain factors are held constant, the main relationships may be restricted. If home background is held constant, that is equivalent to choosing persons all alike in home background. But within so select a group, there would be little variation in intelligence or vocabulary, for example. By holding constant a factor which influences either of the two factors we are comparing, or which is related to either of those factors, we mutilate the correlation. It is just as though we had selected a group all alike in the factors held constant. Finally, all correlations deal only with relatedness and so may often confuse indicators with causes. A more thorough discussion, one worthy of attention by all experimenters using this method, is found in Burks' "On the Inadequacy of the Partial and Multiple Correlation Technique."[1]

How shall we interpret the correlation of -0.35 between religious and ethical scores, when intelligence is constant? It means clearly, that if we select pupils of equal intelligence, there is a distinct tendency for those who make the best scores on a religious test to make scores poorer than average on the ethical test. One tends to interfere with the other, or at least to go with certain kinds of training which interfere with the other. This is not a high negative correlation. There are many exceptions. It is really about 94 per cent pure chance (see Table XI). Yet if these are fair results, there is surely a tendency within this group for something related to what a pupil has learned about religion to be of more harm than good in scoring well on an ethical test. Of course this is a hypothetical case, yet in experiments with several hundred children the writer has found a correlation of 0.23 between religious knowledge and ethical ideas, which, after

[1] *Jour. Edu. Psy.*, Vol. XVII, pp. 532–40, 625–30, November and December, 1926.

intelligence had been partialled out, dropped to 0.02, indicating that the two kinds of training were much less related than is ordinarily supposed.

Uses for Partial Correlation: Because the method of partial correlation opens up so many avenues of experimentation which have seemed closed heretofore, it seems well to give unusual attention to its applications. Not long ago a conference dealing with a national program of religious education was in dispute as to the merits of an all-round program (physical, mental, social, religious, etc.) within each club, as contrasted with a program which offered a series of separate special interest groups. Should young men belong to one omnibus organization, or should they belong to a number of organizations, each serving its separate purpose? Plans for experimentation were laid out, in order to obtain light on the question. It seemed impossible to find leaders who were so much in control of affairs in their communities that they could agree to run one group on one basis and equivalent groups on the other basis. Moreover the variables were hard to control. If different leaders tried different plans, perhaps the apparent difference might be due to the superiority of one of the leaders. Young men might prefer one type of program and resent attempts to develop a different sort. The solution was worked out in a program which involved partial correlations. Let each leader lead each group in the way it chooses to go, but let him keep a careful record so that the *degree* to which the physical, mental, social, and spiritual activities are concentrated within one group can be measured. Some individuals will have a high degree of concentration, others will spread among a number of special interest groups. At the beginning of the work in September, all members were to take a battery of tests, yielding a score along lines felt by the leaders to be significant. In June the tests were to be repeated. Meanwhile the efficiency of each leader was to be rated by state and national officers who supervised his work. A careful record of participation made it clear just how active each individual club member had been. Then the correlation between gains during the year and degree of concentration could be found, with quality of

leadership, and amount of participation, constant. That is, it would be possible to tell whether concentration of activities within a single group would be desirable, undesirable, or indifferent in its influence upon such gains as the tests measured, when every fellow participated just as much as did his neighbor, and when every leader was just as good as the next one. Such information it would be quite impossible to get by any "set-up" experiment.

The above experiment involved holding two items constant, while finding the relationship between a third and fourth. That is, quality of leadership and participation of members were held constant throughout, while the relationship between type of program and results was being measured. The other partial correlations would be hardly less interesting. Thus, what is the relation of type of program to participation, when leadership is constant? What is the relationship of leadership to results when type of program and amount of participation is constant?

It might be possible to use S.D. units and then compare the relative importance of leadership and program, to see upon which one stress should be laid to secure the greatest improvement.

Further illustration of the use of partial correlation in an administrative problem is given on pages 224–237.

f. Partial Regression

Purpose: Just as it was found possible in connection with any correlation to predict with more or less accuracy the score of a given individual in one test or trait, from knowledge of his position in another test or trait, so it is possible to predict by means of partial correlation coefficients from several tests the most probable score of an individual in another test or trait. In the case of the several partial coefficients, prediction is apt to be much more exact. This is particularly true if each of the measures used is independent of each of the others, yet each contributes something toward the factor it is desired to predict.

Of course, in making the prediction not every test receives the same weight that every other test receives. Some are quite useful, others relatively useless. The factor by which each is

multiplied in order to give it its proper weight, is called its regression coefficient.

Computations: For three variables, one of which is predicted by means of scores in the other two, the formula would be as follows:

$$a = W_b (b - Mn_b) + W_c (c - Mn_c) + Mn_a$$

where:

b and c represent the given scores

a is the score to be predicted

W_b, W_c are the weights to be given to b and c in making this prediction

Mn_a, Mn_b, Mn_c are the means for the distributions of a, b, and c.

W_b involves not only the intercorrelations, but the standard deviation also. Its formula is

$$W_b = \frac{r_{ab} - r_{ac}r_{bc}}{1 - r_{bc}^2} \times \frac{\text{S. D.}_a}{\text{S. D.}_b}$$

The S. D.$_a$ and the S. D.$_b$ are, of course, respectively, the standard deviations of a and of b.

In like fashion the weight assigned to c, may be computed:

$$W_c = \frac{r_{ac} - r_{ab}r_{bc}}{1 - r_{bc}^2} \times \frac{\text{S. D.}_a}{\text{S. D.}_c}$$

In the illustration given below, the weights are first computed and then applied to the problem.

Suppose it is desired to predict the ethical score for an individual, about whom it is known only that he made an intelligence score of 142 and a religious score of 98.

From Table IX we know that the S. D. of the intelligence scores is 8.1, that of the ethical scores is 3.7 and that of the religious scores 21.9. The correlations previously figured have told us that

$$r_{er} = 0.13$$
$$r_{et} = 0.66$$
$$r_{rt} = 0.61$$

Hence W_t, the weight for the intelligence scores, is

$$= \frac{r_{et} - r_{er}r_{rt}}{1 - r_{rt}^2} \times \frac{\text{S. D.}_e}{\text{S. D.}_t}$$
$$= \frac{0.66 - 0.13 \times 0.61}{1 - (0.61)^2} \times \frac{3.7}{8.1}$$
$$= 0.43$$

In like manner, W_r the weight for the religious scores is

$$= \frac{r_{er} - r_{et}r_{rt}}{1 - r_{rt}^2} \times \frac{\text{S. D.}^e}{\text{S. D.}_r}$$
$$= \frac{0.13 - 0.66 \times 0.61}{1 - (0.61)^2} \times \frac{3.7}{21.9}$$
$$= -0.07$$

Now that the weights are computed,—and it is interesting to note that the religious scores have a negative weight, indicating that the higher they go, the lower the ethical scores are likely to go,—it is possible to fill in the original equation. The means are, of course, taken from Table IV.

$$e = W_i (i - Mn_i) + W_r (r - Mn_r) + Mn_e$$
$$= 0.43\ (142 - 128.6) - 0.07\ (98 - 63) + 23$$
$$= 5.76 - 2.45 + 23$$
$$= 26.31,$$

the most probable score on the ethical test for any individual who made a score of 142 on the intelligence test and of 98 on the religious test.

When several partial correlations have been computed, and it is desired to combine the scores into a single prediction, it is necessary to compute also the partial standard deviations, which enter into the weighting.

The form of the general regression equation for, let us say, four variables, although longer, is not essentially different from that given above for three. Suppose it is desired to find the most probable score in test A, on the basis of scores in tests B, C, and D. The formula then becomes

$$a = W_b(b - Mn_b) + W_c(c - Mn_c) + W_d(d - Mn_d) + Mn_a$$

One extra term is all that has been added. The real difference has arisen in connection with the weights, W_b, W_c, and W_d. These are a bit more complicated.

$$W_b = r_{ab \cdot cd} \frac{\text{S. D.}_{a \cdot bcd}}{\text{S. D.}_{b \cdot acd}}$$

where

$$\text{S. D.}_{a \cdot bcd} = \text{S. D.}_a \sqrt{(1 - r^2_{ab})\ (1 - r^2_{ac.b})\ (1 - r^2_{ad.bc})}$$

and

$$\text{S. D.}_{b \cdot acd} = \text{S. D.}_b \sqrt{(1 - r^2_{ab})\ (1 - r^2_{bc.a})\ (1 - r^2_{bd.ac})}$$

These equations for the partial standard deviation can be written with the letters following the dot (*i.e.*, those that are held constant) arranged in any order. This means that there are several alternative forms in which the equation can be written, each of which is a check on the arithmetic of the other computation. Frequently by using one form rather than another, it is possible to save computing unnecessary partial correlations.[1]

The weight for each of the other terms may be computed in the same way.

$$W_c = r_{ac \cdot bd} \frac{\text{S. D.}_{a \cdot bcd}}{\text{S. D.}_{c \cdot abd}}$$

[1] The minimum number of equations needed for partial correlation and regression in the case of four or five variables is well illustrated in GARRET, "Statistics in Psychology and Education," pp. 240–251, Longmans, Green, and Co., New York, 1926.

where

$$\text{S. D.}_{c \cdot abd} = \text{S. D.}_{c} \sqrt{(1 - r^2_{ac})\ (1 - r^2_{bc.a})\ (1 - r^2_{dc.ab})}$$

and

$$W_d = r_{ad \cdot bc} \frac{\text{S. D.}_{a \cdot bcd}}{\text{S. D.}_{d \cdot abc}}$$

where

$$\text{S. D.}_{d \cdot abc} = \text{S. D.}_{d} \sqrt{(1 - r^2_{ad})\ (1 - r^2_{bd.a})\ (1 - r^2_{cd.ab})}$$

The procedure for five variables is just like this, except that one more term is involved in the original equation, and one in each of the partial standard deviations. An example is given later.

Interpretation: The weights used in the above equations are often symbolized by b in statistical books, and are called the partial regression coefficients. They indicate the importance of each of the sets of scores in influencing the outcome in the score which is being predicted. They are often used as measures of weight, value, and influence, as well as in connection with the task above outlined, that of predicting an unknown score from certain other scores. Of course the weights assigned are subject to the limitations of all partial correlation results, because of limited samplings, or interrelationships among the variables.

A clear need for this procedure faced a director of Y M C A educational classes. He found that some schools over the country were on a paying basis, whereas others showed an annual deficit. Naturally he wondered why. He set out to see how important each of the following items was in determining surplus or deficit: salary paid faculty per hour of teaching, amount paid by each student per course, number of courses per student, ratio of overhead to number of students, etc. Selecting by inspection of the records those that seemed to be relevant, he found the correlations and partial correlations, and at last built up a regression equation something like this:

Each dollar of gain $= 0.04\ (b - \$6) + 0.23\ (c - \$4.78) - 0.01\ (d - \$.32) - 0.65\ (e - 17 \text{ cents})$

Thus he knew that factor b, if it went above \$6 helped make for gain, but only to 4 per cent of its size. Factor c, however, contributed 23 per cent of whatever it went above \$4.78. Factor d was almost negligible, every dollar above 32 cents making a difference of only 1 cent on the dollar of gain, whereas factor e,

with its regression coefficient of − 0.65 subtracted 65 per cent of every dollar that it rose above 17 cents. Any school administrator could then apply this formula to his situation, discover the sinner with respect to his deficit, and see at just what point changes if he made them, would be most certain to bring about an increase in the surplus.

The contribution of each of several tests to a battery of tests designed, perhaps, for the measure of ethical knowledge, is best determined by the regression coefficient. It may appear that some tests which seem very good so nearly duplicate one another that if one is in, the others make little or no further contribution. This situation, also, is revealed by the regression coefficients.

g. Multiple Correlation

Purpose: Suppose an individual does make out a regression equation and predict from it a series of scores in some other trait. How close will he come? Given tests of intelligence and of religious ideas, is that enough of a basis on which to make a reasonably accurate prediction of ethical score, or will a lot of things be left out of account? Take the case of the Y M C A school administrator. Suppose he should complete his regression equation on the basis of knowledge about amount paid faculty per hour of teaching, of tuition per course, of number of courses per student, and the ratio of overhead expense to the number of students. Would he then have all the important factors, or would there be a multitude of little things, varying from town to town, which would influence the result unpredictably?

It is the purpose of multiple correlation to indicate how closely the desired result can be predicted, with the available measures.

Computation: To predict the best correlation which can be obtained between the scores predicted by two measures, a and b, and those actually resulting c, the formula would be

$$R_{c \cdot ab} = \sqrt{1 - (1 - r_{cb}^2)\,(1 - r_{ca.b}^2)}$$

where:

$R_{c \cdot ab}$ is the multiple correlation possible by the best combination of a and b, in predicting c.

r_{cb} is the correlation between the criterion and one of the factors, b.

$r_{ca \cdot b}$ is the partial correlation between the other factor and the criterion, with the b measures constant.

Let us use the measurements of intelligence and religious score as a means of predicting ethical score. It has been found that the correlation between intelligence and ethical score is 0.66 and that the partial correlation between religious and ethical scores with intelligence constant is − 0.35. Substituting in the above equation we have

$$R_{e \cdot rI} = \sqrt{1 - (1 - 0.66^2)(1 - (-0.35^2))}$$
$$= \sqrt{1 - (0.56 \times 0.88)}$$
$$= 0.71$$

That is to say, by combining a religious test score with an intelligence test score, in the proper proportions we could predict a series of ethical scores that would correlate 0.71 with the actual ones. This is not high. Apparently ethical judgment involves many factors other than those which are measured by intelligence tests, or by religious tests. Intelligence scores alone would predict ethical scores with a correlation of 0.66 so the added knowledge of how a pupil did on his religious test would help very little.

The optimum weight to give the highest correlation depends on the intercorrelations and the standard deviations of each series.

For the case in which five test scores, say b, c, d, e, and f, are all available, to use in predicting the score in trait a, the formula becomes:

$$R_{a(bcdef)} = 1 - \sqrt{(1 - r^2_{ab})\,(1 - r^2_{ac.b})\,(1 - r^2_{ad.bc})\,(1 - r^2_{ae.bcd})\,(1 - r^2_{af.bc})}$$

Interpretation: The higher the coefficient of multiple correlation the more accurate will be the predictions of the regression equation. When the multiple correlation coefficient is low, it means that not enough tests have yet been applied to predict with any degree of accuracy the criterion against which prediction is being applied. The experimenter should search for tests which will correlate high with his criterion, but low with any tests he already has included. Such tests will supplement his present battery and improve his prediction. The percentage table (Table XXXII) for the interpretation of correlations is very useful here, as is also the interpretation in terms of displacement. (Table XXXI) A multiple-correlation coefficient of 0.95 means that 75 per cent of the factors tending to produce the result in the case of the criterion, will have been measured and taken properly into account by the regression equation. To predict individual scores with a negligible margin of error, it is desirable to have multiple correlations of 0.99 and above.

7. A Practical Problem in Five Variables

Partial correlation, regression, and multiple correlation work together in the solution of many involved problems. They open so many doors to the experimenter, that a hypothetical problem is here presented.

Suppose a denominational board in charge of religious education is faced each year with the problem of determining what emphasis it shall place upon various methods of improving church-school work. Shall it stress the adoption of improved lesson materials? Shall it stress more training for teachers? Shall it promote methods of follow-up which will prevent absences and tardiness? Shall it try to get more adequate plants and equipment for religious education? There may be a feeling that all of these should be done. Then, in what proportion? Is one as important as another? What combination will produce the best return for a given expenditure of time and money?

Several types of experimental method might be employed to answer such questions. Equated fields might be chosen, and one method stressed in one field, another method in a field of equal difficulty and promise elsewhere. The difficulty would be to equate the fields and to keep the stress on every other element negligible. The procedure here suggested is one of the survey type, in which the experimental conditions are set up statistically, the children being taken in their natural setting. In practice, the results of such a study should be tested and confirmed by actual experiment, to distinguish real causes from related but really ineffective symptoms. Suppose that an endeavor was made to test all children in October, 1930. Suppose that during the year careful record was kept of expenditures, and that in October, 1931, the children were retested. Suppose the following additional information to have been secured.

1. Type of lesson material used. This could later be graded on a score card, ranging perhaps from zero to 20, with 10 as average. Samples could be sent in by classes not using standard courses.

2. Training of the teacher. This would involve years of schooling, and the amount of time spent in the special study of religious education, in institutes, teacher training classes, schools of religious education, etc. The various elements should be combined so as to yield one composite score for the training of the teacher. Later this can be analyzed more carefully, into its various types. At present the aim is to separate teachers well trained from those less trained. Suppose a 20-point scale be used for this, also.
3. The physical equipment, building, books, classrooms, recreational facilities, etc. Perhaps a questionnaire might gather such items, perhaps they could better be obtained by scoring each building on the basis of some such score card as that suggested in Volume II of the "Indiana Survey of Religious Education." Suppose the buildings were scored on a scale giving 250 possible points.
4. The number of times each pupil has been absent during the year. Tardiness may be counted as a fraction of an absence.

Suppose reliable information was forthcoming from 500 church schools scattered throughout the country, some city and some rural. Suppose that the answers represented about 100 children per school, for whom there were two test scores, and the above information. It would then be possible to find for each pupil his gain in the sort of thing measured by the test or battery of tests.

Suppose the following table to represent the average, and the standard deviation for each item. After each is stated the cost, during the year 1930–1931, necessary to secure a movement of one point on each scale for the average church school. Thus to bring about a rise of one step in the average test score, cost during the year $28.25. Other experiments had indicated, it is assumed, that to raise the average lesson materials one point on the score card, cost the denominational board $2.05 per school; to improve teacher training one point cost $5.27 per school, etc. It may be remarked that such a system of accounting is not yet in evidence in most denominational boards, but it is not at all impracticable.

TABLE XXI. — SUMMARY OF DATA FOR HYPOTHETICAL PROBLEM IN ADMINISTRATION OF BOARD OF RELIGIOUS EDUCATION

Item	Symbol used in formulae	Average	Standard deviation	Cost to move average church school one point on scale
1. Gain in test scores	*G*	33	10	$28.25
2. Lesson materials in use	*L*	12	5	2.05
3. Teacher training	*T*	8.8	3.2	5.27
4. Plant and physical equipment	*P*	126	60	0.12
5. Absences	*A*	28	12	0.54

Suppose the gain in score for each pupil on the tests was correlated with every other item, and that lesson materials be correlated with the remaining items, teacher-training scores with plant and absence, and scores for plant, with absence scores. We will then have correlations between each item and every other. Suppose them to have been something as follows. This is a purely hypothetical case, but the endeavor has been made to keep the figures in accord with good judgment as to what would probably be found.

TABLE XXII. — INTERCORRELATIONS OF MEASURES TAKEN BY BOARD

	L	*T*	*P*	*A*
G	0.62	0.85	0.25	− 0.06
L	....	0.40	0.30	− 0.50
T	....	....	0.35	− 0.20
P	....	....	....	− 0.45

On the face of them, these correlations indicate that large gains are made by pupils with highly trained teachers, and, on the whole, by pupils using excellent lesson material. There is some tendency for pupils absent a great deal to make low gains, and

for fair gain to be made where there is good physical plant. The last does not appear to be so closely related to gain. Looking at the T row we observe that where teachers are well trained there tends to be good lesson material and a good physical plant, but this is not invariable at all. Also there tend to be fewer absences. Absence seems to correlate negatively with all other items, indicating that where teaching, lesson materials, and physical plant are high, amount of absence is low, and *vice versa*.

Applying partial correlation, let us first eliminate the variations due to differences in teacher training. Bear in mind that as this is done some of the other relationships will be distorted because they are not equally related to teacher training. The formula for each partial correlation will be similar to that given on page 214 and illustrated in Table XX.

$$r_{GL \cdot T} = \frac{r_{GL} - (r_{GT})(r_{LT})}{\sqrt{(1 - r_{GT}^2)(1 - r_{LT}^2)}} = \frac{0.62 - (0.85 \times 0.40)}{\sqrt{(1 - 0.85^2)(1 - 0.40^2)}} = \frac{0.28}{\sqrt{0.28 \times 0.84}} = 0.57 \qquad (1)$$

$$r_{GA \cdot T} = \frac{r_{GA} - (r_{GT})(r_{AT})}{\sqrt{(1 - r_{GT}^2)(1 - r_{AT}^2)}} = \frac{-0.60 - (0.85 \times -0.20)}{\sqrt{(1 - 0.85^2)(1 - [-0.20]^2)}} = \frac{-0.43}{\sqrt{0.28 \times 0.96}} = -0.83 \qquad (2)$$

$$r_{LA \cdot T} = \frac{r_{LA} - (r_{LT})(r_{AT})}{\sqrt{(1 - r_{LT}^2)(1 - r_{AT}^2)}} = \frac{-0.50 - (0.40 \times -0.20)}{\sqrt{(1 - 0.40^2)(1 - [-0.20]^2)}} = \frac{-0.42}{\sqrt{0.84 \times 0.96}} = -0.47 \qquad (3)$$

$$r_{GP \cdot T} = \frac{r_{GP} - (r_{GT})(r_{PT})}{\sqrt{(1 - r_{GT}^2)(1 - r_{PT}^2)}} = \frac{0.25 - (0.85 \times 0.35)}{\sqrt{(1 - 0.85^2)(1 - 0.35^2)}} = \frac{-0.05}{\sqrt{0.28 \times 0.28}} = -0.10 \qquad (4)$$

$$r_{LP \cdot T} = \frac{r_{LP} - (r_{LT})(r_{PT})}{\sqrt{(1 - r_{LT}^2)(1 - r_{PT}^2)}} = \frac{0.30 - (0.40 \times 0.35)}{\sqrt{(1 - 0.40^2)(1 - 0.35^2)}} = \frac{0.16}{\sqrt{0.84 \times 0.88}} = 0.19 \qquad (5)$$

$$r_{AP\cdot T} = \frac{r_{AP} - (r_{AT})(r_{PT})}{\sqrt{(1 - r_{AT}^2)(1 - r_{TP}^2)}} = \frac{-0.45 - (-0.20 \times 0.35)}{\sqrt{(1 - [-0.20]^2)(1 - 0.35^2)}}$$

$$= \frac{-0.38}{\sqrt{0.96 \times 0.88}} = -0.41 \qquad (6)$$

These equations show what the relationships among the other factors would have been had all teachers been of equal training. Equation (1) yielding a correlation of 0.57 shows that there is a slight decrease in the relationship between the gain and the value of lesson material, when the teacher's training has been held constant. Some of what looked like gain due to lessons, may have been due to superior teachers who used the lessons. Absences are shown by equation (2) to give a high negative correlation with gain. That is, when teaching ability is the same, the better the lesson material the fewer the absences. This is more apparent than it was when teacher training was uncontrolled, and confused the relationship. Equation (4) indicates that if this factor of differences in teaching ability is held constant, the physical plant seems to be far less significant than it had appeared to be from the first correlations, perhaps indeed tending to be a hindrance.

In a thoroughgoing investigation, it would be wise to find the first order partials for all variables. That is, the results could be studied, not only with teaching preparation constant, but then with absences, physical plant, lesson materials, each in turn held constant. Equations (12) to (17) present the results when lesson materials are held constant. At present let us finish the study of the relationship of lesson materials to gain, when other elements are constant. The next step is the separating out of the influence exerted by physical plant. These second order partials build upon the first six equations, so that in the results given below, both teacher training and physical plant are held constant.

$$r_{GL\cdot TP} = \frac{r_{GL\cdot T} - (r_{GP\cdot T})(r_{LP\cdot T})}{\sqrt{(1 - r_{GP\cdot T}^2)(1 - r_{LP\cdot T}^2)}} = \frac{0.57 - (-0.10 \times 0.19)}{\sqrt{(1 - [-0.10]^2)(1 - 0.19^2)}} =$$

$$\frac{0.59}{\sqrt{0.99 \times 0.96}} = 0.60 \qquad (7)$$

$$r_{GA\cdot TP} = \frac{r_{GA\cdot T} - (r_{GP\cdot T})(r_{AP\cdot T})}{\sqrt{(1-r_{GP\cdot T}^2)(1-r_{AP\cdot T}^2)}} = \frac{-0.83-(-0.10\times-0.41)}{\sqrt{(1-[-0.10]^2)(1-[-0.41]^2)}}$$

$$= \frac{-0.87}{\sqrt{0.99\times0.83}} = -0.97 \qquad (8)$$

$$r_{LA\cdot TP} = \frac{r_{LA\cdot T} - (r_{LP\cdot T})(r_{AP\cdot T})}{\sqrt{(1-r_{LLP\cdot T}^2)(1-r_{AP\cdot T}^2)}} = \frac{-0.47-(0.19\times-0.41)}{\sqrt{(1-0.19^2)(1-[-0.41]^2)}}$$

$$= \frac{-0.39}{\sqrt{0.96\times0.83}} = -0.44 \qquad (9)$$

These three equations indicate that when not only teacher training but building equipment are controlled, gain due to lesson materials is not much influenced, the absences become much more important as determiners of gain. Equation (9) being very much like equation (3) and like the original correlation of − 0.50 shows the relationship between type of lesson and amount of absence little influenced by differences between schools in teacher training and in physical equipment. Using the percentage interpretation (Table XXXI) it may be said that bad lesson material and absence from Sunday school have one-third of their causes in common, and this holds even when teaching and building environment have been fairly well equated. Of course, poor material does not by itself cause one-third of the absence, because both tend to be produced by low-grade uninterested constituencies. The fact, however, that the per cent of common cause dropped so little when teacher training and type of building, both pretty fair indices of constituency, have been held constant, indicates that probably lesson material is at least a 20 per cent cause of absences.

The final step, toward which the first nine equations were preliminary operations, is the determination of the correlation between gain as shown by the tests, and lesson material, when teacher training, physical equipment, and absence have all been held constant.

$$r_{GL\cdot TPA} = \frac{r_{GL\cdot TP} - (r_{GA\cdot TP})(r_{LA\cdot TP})}{\sqrt{(1-r_{GA\cdot TP}^2)(1-r_{LA\cdot TP}^2)}}$$

$$= \frac{0.60-(-0.97\times-0.44)}{\sqrt{(1-[-0.97]^2)(1-[-0.44]^2)}} = 0.78 \qquad (10)$$

There is thus a significant tendency for pupils in classes which use better lesson materials to make larger gains, even though the classes have teachers who are alike in training, meet in the same sort of physical environment, and show no differences in attendance. In fact, the influence of this lesson material on gain stands out more clearly when the other possible influences have been partialled out. It is not possible from these data to be sure that lesson material was the cause of the gain, even to the extent indicated by a correlation of 0.78, for certain other factors were not measured. Thus communities with more intelligent parents, higher home standards, better public schools, and better playgrounds, also may have better curricula, and what looks like an influence due to curricula may be due to one or another of these causes. In so far as the type of church building is a measure of those things, and in so far as the type of teachers is an index to community standard, they have already been partialled out. It may be that more thorough-going investigation would still indicate that gain is closely related to the type of curriculum material used.

With the first nine equations it is possible to figure a second final equation.

$$r_{GA\cdot LTP} = \frac{r_{GA\cdot TP} - (r_{GL\cdot TP})(r_{AL\cdot TP})}{\sqrt{(1 - r_{GL}{}^2{}_{\cdot TP})(1 - r_{AL}{}^2{}_{\cdot TP})}} = \frac{-0.97 - (0.60 \times -0.44)}{(1 - 0.60^2)(1 - [-0.44]^2)}$$

$$= -0.97 \quad \textbf{(11)}$$

Here we seem to have a factor that is related in first class fashion to the kind of gain which the tests measure. If all pupils had the same sort of teachers, the same materials, the same building, then absence from church school would be an excellent measure of their lack of achievement. The high negative correlation indicates that those who are absent least almost invariably make the highest scores, and that those absent most almost invariably fall low. Turning to Table XXXI for the interpretation of correlation, we find that if the lesson materials, teacher training, and plant are a constant, as they are apt to be in any given class, 80 per cent of the other causes of gain are related to or summed up in this matter of absence. The corre-

lation of − 0.60 with gain with which we started did not make it at all clear that absence played so important a rôle. Of course, absence may be the indication of other things beside itself that really influence score. It may mean lack of interest, lack of home coöperation, etc. On the basis of these data, however, it is an excellent index.

To study the relationship of gain to the other factors, it will be necessary to start over again, working through a series of six first-order partials, and three second-order partials. The formulae will not be repeated here, but can be inferred from the coefficients used.

$$r_{GT \cdot L} = \frac{0.85 - (0.62 \times 0.40)}{\sqrt{(1 - 0.62^2)(1 - 0.40^2)}} = 0.82 \qquad (12)$$

$$r_{GP \cdot L} = \frac{0.25 - (0.62 \times 0.30)}{\sqrt{(1 - 0.62^2)(1 - 0.30^2)}} = 0.08 \qquad (13)$$

$$r_{TP \cdot L} = \frac{0.35 - (0.40 \times 0.30)}{\sqrt{(1 - 0.40^2)(1 - 0.30^2)}} = 0.26 \qquad (14)$$

$$r_{GA \cdot L} = \frac{-0.60 - (0.62 \times -0.50)}{\sqrt{(1 - 0.62^2)(1 - [-0.50]^2)}} = -0.42 \qquad (15)$$

$$r_{TA \cdot L} = \frac{-0.20 - (-0.50 \times 0.40)}{\sqrt{(1 - [-0.50]^2)(1 - 0.40^2)}} = 0.00 \qquad (16)$$

$$r_{PA \cdot L} = \frac{-0.45 - (-0.50 \times 0.30)}{\sqrt{(1 - [-0.50]^2)(1 - 0.30^2)}} = -0.36 \qquad (17)$$

These reveal that gain is related to teaching preparation in about the same degree whether teaching materials vary (0.85) or are held constant (0.82). There seems to be very little relationship between gain and physical plant (0.08) when lesson material is constant. Oddly enough, if all classes using the same materials had been selected, the correlation between teaching and absence, instead of being − 0.20 would become zero. This suggests that attendance of pupils may be more closely related to other factors—the physical equipment, lesson materials, etc.—than it is to teacher preparation. Further study should separate this out more distinctly.

Now the second-order partials will be built up by eliminating from the first-order group just completed the additional factor of absence.

$$r_{GT \cdot LA} = \frac{0.82 - (-0.42 \times 0.00)}{\sqrt{(1 - [-0.42]^2)(1 - 0.00^2)}} = 0.90 \qquad (18)$$

$$r_{GP \cdot LA} = \frac{0.08 - (-0.42 \times -0.36)}{\sqrt{(1 - [-0.42]^2)(1 - [-0.36]^2)}} = -0.08 \qquad (19)$$

$$r_{TP \cdot LA} = \frac{0.26 - (0.00 \times -0.36)}{\sqrt{(1 - 0.00^2)(1 - [-0.36]^2)}} = 0.28 \qquad (20)$$

These indicate that if we had all pupils with the same lesson materials, and the same number of absences, their gain would be very closely related to the training of their teachers. It would be slightly influenced by plant, and that in a negative direction (− 0.08). Even in churches with the same grade of lesson materials and the same attendance records, there is a tendency for those with better plants to have better trained teachers, and *vice versa*, but this relationship is not large (0.28).

The third-order partials will give us the correlation of gain with teacher training and plant, when freed from the influence of the other variables.

$$r_{GT \cdot LAP} = \frac{0.90 - (-0.08 \times 0.28)}{\sqrt{(1 - [-0.08]^2)(1 - 0.28^2)}} = 0.97 \qquad (21)$$

$$r_{GP \cdot LAT} = \frac{-0.08 - (0.90 \times 0.28)}{\sqrt{(1 - 0.90^2)(1 - 0.28^2)}} = -0.79 \qquad (22)$$

It appears that, given pupils all of whom have studied the same material, all of whom have been present an equal number of times, and all of whom have the advantages of the same equipment, the differences remaining will correlate almost perfectly (0.97) with the training their teachers have had. There are few factors influencing gain not represented by this. This is probably not so surprising as the other coefficient (− 0.79) which suggests that if pupils have the same teachers, materials, and attendance, the building is correlated with failure rather than with success. This may be interpreted on the basis of schools with good equipment for recreation, etc. attracting persons less interested in the

sort of thing measured by the test, or at least persons from homes where less interest is shown than is shown in those homes represented in church schools which, in spite of poor equipment, have managed to keep lessons, teachers, absences, etc., on a good plane. At any rate, such a figure if found in an actual investigation would raise serious questions about the function of equipment.

The next question the board might care to raise would be, "Have we really gathered in all the evidence? Do these four factors really represent the things making for success in the tests, or are there many variables which, if we use only these, we shall miss?" The answer is found in the multiple correlation.

(23)

$$
\begin{aligned}
R_{G\cdot TPAL} &= 1-\sqrt{(1-r^2_{GT})(1-r^2_{GP\cdot T})(1-r^2_{GA\cdot PT})(1-r^2_{GL\cdot APT})} \\
&= 1-\sqrt{(1-0.85^2)(1-[-0.10]^2)(1-[-0.97]^2)(1-0.78^2)} \\
&= 1-\sqrt{0.28\times 0.99\times 0.06\times 0.39} \\
&= 1-\sqrt{0.00649} \\
&= 0.92
\end{aligned}
$$

It appears that if these four measured factors about children in church schools be weighted in the most favorable fashion, they will yield a correlation of 0.92 with the gains which take place on test results. Using Table XXXII this may be interpreted as meaning that these measures cover about 70 per cent of the factors at work to produce differences in gain among the pupils in the Sunday schools measured; 30 per cent are not included by these measures. Still, it is probable that they correlate with test results as well as test results would correlate with themselves, for tests with reliabilities above 0.92 are rather rare.

Since the board has within its control, so far as it can influence lesson materials, teacher training, absence, and buildings, the major factors related to gains during the year in church schools, how important is each? Just what is the most favorable weight to be assigned to each of these factors? If a combination attack is to be made, what is the optimum distribution of time and money? What experiments are most worth trying?

The next step toward answering such questions, is the computation of the weights to be used in the regression equation.

Those, it will be remembered, involve not only partial correlation, but also certain partial standard deviations. The influence a trait has depends in part upon the range of its scores, other things being constant. In order to build the regression equation for gain, five partial standard deviations are needed.

$$\text{S.D.}^{**}_{G\cdot LATP} \tag{24}$$
$$= \text{S.D.}^{*}_{G}\sqrt{(1-r^2_{GL})(1-r^2_{GA\cdot L})(1-r^2_{GT\cdot AL})(1-r^2_{GP\cdot TAL})}$$
$$= 10\sqrt{(1-0.62^2)(1-[-0.42]^2)(1-0.90^2)(1-[-0.79]^2)}$$
$$= 10\sqrt{0.62\times 0.82\times 0.19\times 0.37}$$
$$= 1.88$$

$$\text{S.D.}_{L\cdot TPAG} \tag{25}$$
$$= \text{S.D.}_{L}\sqrt{(1-r^2_{LT})(1-r^2_{LP\cdot T})(1-r^2_{LA\cdot PT})(1-r^2_{LG\cdot APT})}$$
$$= 5\sqrt{(1-0.40^2)(1-0.19^2)(1-[-0.44]^2)(1-0.78^2)}$$
$$= 5\sqrt{0.84\times 0.96\times 0.81\times 0.49}$$
$$= 2.81$$

$$\text{S.D.}_{A\cdot TPLG} \tag{26}$$
$$= \text{S.D.}_{A}\sqrt{(1-r^2_{AT})(1-r^2_{AP\cdot T})(1-r^2_{AL\cdot PT})(1-r^2_{AG\cdot LPT})}$$
$$= 12\sqrt{(1-(-0.20)^2)(1-(-0.41)^2)(1-(-0.44)^2)(1-(-0.97)^2)}$$
$$= 12\sqrt{0.96\times 0.83\times 0.81\times 0.06}$$
$$= 2.36$$

$$\text{S.D.}_{T\cdot LAPG} \tag{27}$$
$$= \text{S.D.}_{T}\sqrt{(1-r^2_{TL})(1-r^2_{TA\cdot L})(1-r^2_{TP\cdot AL})(1-r^2_{TG\cdot PAL})}$$
$$= 3.2\sqrt{(1-0.40^2)(1-0.00^2)(1-0.28^2)(1-0.97^2)}$$
$$= 3.2\sqrt{0.84\times 1.00\times 0.92\times 0.06}$$
$$= 0.69$$

$$\text{S.D.}_{P\cdot LATG} \tag{28}$$
$$= \text{S.D.}_{P}\sqrt{(1-r^2_{PL})(1-r^2_{PA\cdot L})(1-r^2_{PT\cdot AL})(1-r^2_{PG\cdot TAL})}$$
$$= 60\sqrt{(1-0.30^2)(1-(-0.36)^2)(1-0.28^2)(1-(-0.79)^2)}$$
$$= 60\sqrt{0.91\times 0.87\times 0.92\times 0.38}$$
$$= 31.57$$

* Standard deviations given in Table No. XXI.

** It makes no difference in what order the subscripts following the dot appear. They are here arranged to make use of correlations already computed.

It is now possible to find the proper weight to be assigned to each of the four factors when they are to be used in predicting gain.

1. Weight for Lesson Materials . . . $W_L = r_{GL \cdot ATP} \frac{S.D._{G \cdot LATP}}{S.D._{L \cdot GATP}}$

$$= 0.78 \times \frac{1.88}{2.81}$$

$$= 0.52$$

2. Weight for Absences . . . . . $W_A = r_{GA \cdot LTP} \frac{S.D._{G \cdot LATP}}{S.D._{A \cdot GLTP}}$

$$= -0.97 \times \frac{1.88}{2.36}$$

$$= -0.77$$

3. Weight for Teacher Training . . . $W_T = r_{GT \cdot PAL} \frac{S.D._{G \cdot LATP}}{S.D._{T \cdot GPAL}}$

$$= 0.97 \times \frac{1.88}{0.69}$$

$$= 2.64$$

4. Weight for Physical Equipment . . $W_P = r_{GP \cdot ALT} \frac{S.D._{G \cdot LATP}}{S.D._{P \cdot GALT}}$

$$= -0.79 \times \frac{1.88}{31.57}$$

$$= -0.05$$

This tells us that for every rise of one step in gain score, there has tended to be a rise of 0.52 of a score in the value assigned to lesson materials, a decrease of 0.77 of a time in the number of absences, a rise of 2.64 points in the training of teachers, and a decrease of 0.05 of a point in the score assigned to the equipment. These are not yet strictly comparable because one unit on the building scale is not at all the same amount of progress as one unit on the absence scale or the teacher training scale. To judge relative importance it will be necessary to use some comparable measure for each of these score units. This might be done in terms of percentile or standard deviation units. Inasmuch as we are dealing with the problem of an administrative board, it may well be done in terms of costs.

Table XXI showed that to bring about a rise of one point in

the teacher training score cost $5.27. The lesson materials cost $2.05 per point, to prevent one absence cost 54 cents, and to bring about a gain of one point of the building point card on the average Sunday school cost only 12 cents. These figures, it has been assumed, have been determined on the basis of the study on the costs and results during previous years. Using these figures which tell us the cost to the board of a rise of one unit above the average, we may discover the cost to bring about a rise of one point in the average gain. The regression equation would be:

$$G = 0.52\,L + 0.77\,A + 2.64\,T + 0.05\,P$$

Positive signs have been used in the equation all the way through because it is assumed that it is understood that the expenditure will be in the direction of decreasing absences and of discouraging building equipment. Substituting the cost values we have cost of one step of gain in test score in the average Sunday school
= 0.52 ($2.05) + 0.77 ($0.54) + 2.64 ($5.27) + 0.05 ($0.12)
= $1.07 to be expended upon the promotion of better lesson materials.
+ $0.39 to be expended in trying to prevent absences.
+ $13.91 to be expended in the promotion of teacher training.
+ $0.01 to be expended in discouraging building programs.
= $15.38

This is based upon the cost of one step of gain in score on the tests with which progress has been measured. If it is desired to interpret it in terms of the way in which each dollar of money available should be divided in order to secure the largest progress, then we may divide each figure by $15.38, obtaining:

For Lesson Materials	$0.07
For Absence Prevention	0.025
For Teacher Training	0.90
For Building Prevention	0.005
	$1.00

These results indicate[1] that if conditions stay the same as they were last year, and if a dollar at one place on the scale accom-

[1] See Table XXV, p. 250.

plishes as much as a dollar at any other place,[1] then in order to get maximum gain, they may well spend fourteen times as much upon increasing the extent of teacher training as upon improving lesson materials, and thirty or forty times as much upon teacher training as upon devices and campaigns for increasing attendance. From the standpoint of the gains in which the board is interested, building and equipment may well receive no attention at all. Actual experiment would be needed to prove the validity of these apparent causes.

Such an investigation, covering 50,000 children, would cost a great deal of money for test preparation, administration, scoring, tabulation, and statistical labor. It might cost ten or twenty thousand dollars, but this is an insignificant part of what it would save. The figures given in Table XXI showed that $466,125 was spent during the year 1930–1931. That meant that an average gain of 33 points in each of 500 schools cost $28.25 per point. If the money were spent in accord with the emphasis suggested in the regression equation, each point of gain would cost only $15.38 and to bring about such gain as took place in the previous year would cost only $253,770, a saving of $212,355. The expenditure of $10,000 or $20,000 for research in such a situation seems to be amply justified.

Measures of Reliability

Purpose: Every measurement is affected by a variety of errors. This is quite as true of physical measurements as of mental and social measurements. Indeed it is probable that many physical diagnoses are far less reliable than the diagnoses of a psychologist with reference to a child's mental ability. One of the common errors which affects any experiment is due to the fact that the sampling is not perfect. The above experiment may have been carried on with 500 church schools, but 500 church schools are not all the church schools in a given denomination, and if a second 500 were studied, the results obtained might vary slightly from the first 500 measured. Measures of reliability

[1] This is obviously not the case. The dollars spent near the beginning of the work might be more productive than those spent after excellent results had already been achieved. Returns could hardly continue indefinitely at the same rate.

(or, perhaps, measures of "unreliability") tell how much variation may be expected from chance fluctuations, when other samples are chosen.

a. S.D. and *P.E.*

There are two measures of reliability in common use. The first is the *standard deviation*, sometimes in this connection called the standard error. The second is the *probable error* (P.E.) which is always 0.6745 times the S.D. While the S.D. is a distance along the base line of a normal curve which includes about 68 per cent of the cases, P.E. is a shorter distance including just half of the cases. The standard deviation may be found for any measurement or statistical result. From this the probable error may be computed. Unless it is particularly desired to state the results in terms of probable error, the standard deviation may better be used.

Computation and Interpretation: The chances are two to one that the true measure for a very large number of cases is within 1 S.D. of the obtained result. The chances are 21 to 1 that the true measure is within 2 S.D.'s of the obtained measure. The chances are 369 to 1, or practical certainty, that the true measure is within 3 S.D. measures. These ratios are applicable to every reliability measure.

The standard deviation of a mean is equal to the standard deviation of the distribution divided by the square root of the number of cases in the distribution. This is illustrated in Table IX where the standard deviation for the mean of the intelligence scores, ethical scores, and religious scores has been computed. Each of these mean averages was based upon only 25 cases. Suppose there had been 10,000 cases of which the 25 were only a fair sample, it is practically certain that the mean for the 10,000 cases would have fallen within three standard deviations either side of the mean. That is, the mean obtained for the intelligence scores was 128.6. Since the standard deviation of that mean is 1.62, there is practical certainty that had 10,000 pupils been tested, the true mean would have fallen not higher than 133.6, and not lower than 123.6, this representing a distance of 3 S.D.'s either side of the obtained mean. In the case of the ethical

scores the obtained mean was 24, and the standard deviation for that mean 0.74. We can therefore be practically certain that the true mean for any group of which this group was a chance sample would not be higher than 26.2 nor lower than 21.8.

In the case of the median, its standard deviation as illustrated in Table IX is found by multiplying the original standard deviation by 1.25 and then dividing by the square root of the number of cases. The standard deviation of a standard deviation is found in just the way in which the standard deviation of the mean is found, except that it is necessary to divide by the square root of $2n$ instead of just n, where n is the number of cases. The S. D. of Q is found by taking $1.11 \times$ S. D. and dividing the result by the square root of $2n$. The S. D. of a correlation is equal to one minus the correlation squared, divided by the square root of the number of cases. The formula on page 240 shows the probable error for a correlation of any given size based upon certain numbers of cases. It then appears that a correlation of 0.61 based upon 25 cases has a probable error of 0.126. This can be converted into standard deviation, of course, by dividing by 0.6745. Or it may, if desired, be interpreted in terms of practical certainty by using a range of 4.4 probable errors. The chances are 369 to 1 that the true measure is within 4.4 probable errors of the obtained measure. This method for finding the probable error or standard deviation of a correlation is applicable also to partial correlations as is suggested by the illustration in Table XXIII.

TABLE XXIII. — THE RELIABILITY OF CERTAIN MEASURES PREVIOUSLY COMPUTED

I. Means:

Intelligence:

$$\text{S. D.}_{Mn} = \frac{\text{S. D.}}{\sqrt{n}} = \frac{8.1}{5} = 1.62$$

$$\text{P. E.}_{Mn} = 0.6745 \times 1.62 = 1.09$$

Ethical:

$$\text{S. D.}_{Mn} = \frac{\text{S. D.}}{\sqrt{n}} = \frac{3.7}{5} = 0.74$$

$$\text{P. E.}_{Mn} = 0.6745 \times .74 = 0.50$$

Religious:

$$\text{S. D.}_{Mn} = \frac{\text{S. D.}}{\sqrt{n}} = \frac{21.9}{5} = 4.38$$

$$\text{P. E.}_{Mn} = 0.6745 \times 4.38 = 2.95$$

II. Medians:

Intelligence:

$$\text{S. D.}_{M} = \frac{\text{S. D.} \times 1.25}{\sqrt{n}} = \frac{10.13}{5} = 2.03$$

$$\text{P. E.}_{M} = 0.6745 \times 2.03 = 1.37$$

III. Standard Deviation:

Intelligence:

$$\text{S. D.}_{\text{S. D.}} = \frac{\text{S. D.}}{\sqrt{2n}} = \frac{8.1}{7.07} = 1.14$$

$$\text{P. E.}_{\text{S. D.}} = 0.6745 \times 1.14 = 0.77$$

IV. Q:

Intelligence:

$$\text{S. D.}_{Q} = \frac{\text{S. D.} \times 1.11}{\sqrt{2n}} = \frac{8.99}{7.07} = 1.27$$

$$\text{P. E.}_{Q} = 0.6745 \times 1.27 = 0.86$$

V. Correlations:

Intelligence and Religious Score:

$$\text{S. D.}_{r_{ir}} = \frac{1 - r^2}{\sqrt{n}} = \frac{1 - 0.61^2}{\sqrt{25}} \frac{0.63}{5} = 0.126$$

$$\text{P. E.}_{r_{ir}} = 0.6745 \times 0.126 = 0.085$$

Religious and Ethical Score, with Intelligence Constant:

$$\text{S. D.}_{r_{er.i}} = \frac{1 - r^2}{\sqrt{n}} = \frac{1 - (-0.35)^2}{\sqrt{25}} = \frac{0.88}{5} = 0.176$$

$$\text{P. E.}_{r_{er.i}} = 0.6745 \times 0.176 = 0.12$$

b. Reliability of Prediction from a Regression Equation

The standard deviation of a score estimated on the basis of a regression equation may be found from the following simple formula:

$$\text{S. D. predicted score in } A = \text{S. D.}_{A} \sqrt{1 - r_{AB}^2}$$

Thus, in the case of the prediction of ethical score on the basis of intelligence, it was predicted in Table XIX that the score would be 23.33. In Table XXIV it is shown that the S. D. of that predicted score is 2.75, which means that the true value of the score can be, with reasonable certainty, predicted to lie between 15.08 and 31.58, a range wider than the total range within the class, hence not a very useful prediction. Table XXIV also shows that the S. D. of the ethical score predicted on the basis of religious test score is 3.66. Since the prediction was 22.59 it is practically certain that the true value for the ethical score in this case lies between 11.61 and 33.56, a still more useless prediction. Whenever a regression equation is used for prediction, the S. D. or the P. E. of the prediction should be stated. Clearly the

higher the correlation between the traits the less significant is this error of prediction. The S. D. of the prediction is large enough to make prediction of very little use unless the correlations are above 0.90.

TABLE XXIV. — THE RELIABILITY OF THE PREDICTIONS IN TABLE XIX

S. D. Predicted ethical score from given intelligence	$= \text{S. D.}_e\sqrt{1 - r_{ei}^2}$
	$= 3.7\sqrt{1 - 0.66}$
	$= 3.7 \times 0.75$
	$= 2.75$
S. D. Predicted ethical score from given religious score	$= \text{S. D.}_e\, 1 - r_{er}^2$
	$= 3.7 \quad 1 - .13^2$
	$= 3.7 \times 0.99$
	$= 3.66$

c. Reliability of a Difference

The question of reliability is extremely important in experiments in which the object is to find the difference between the result produced by one operation and the result produced by some other operation. Suppose, for example, that it is found that one textbook leads to a gain of 73 whereas another textbook leads to an average gain of 78. Is it clear that the second is the better book? It is not necessarily true. It depends in part upon the probable error or standard deviation of this difference of 5. The standard deviation of a difference, one of the most important measures in all statistics, is approximately[1] equal to the square root of the sum of the squares of the standard deviations of the measures. Thus,

$$\text{S. D.}_{\text{Difference}} = \sqrt{\text{S. D.}^2_{\text{Measure 1}} + \text{S. D.}^2_{\text{Measure 2}}}$$

Suppose that in this given example the standard deviation of the

[1] This is based upon the assumption that there is no correlation between the two series of measures being compared. If the correlation is significant then the full formula should be used.

$$\text{S. D.}_{\text{Difference}} = \sqrt{\text{S. D.}^2_{\text{Measure 1}} + \text{S. D.}^2_{\text{Measure 2}} - 2r_{12}\,(\text{S. D.}_1)\,(\text{S. D.}_2)}$$

The error in the abbreviation is on the side of caution.

mean in one case was 11 and in the other case was 14, the standard deviation of the difference would be

$$\begin{aligned} &\sqrt{11^2 + 14^2} \\ &= \sqrt{121 + 196} \\ &= \sqrt{317} \\ &= 17.9 \end{aligned}$$

This means that a difference to be surely significant in such a case would have to be 3×17.9 or 54. If the difference in gain were less than 54, no one could be certain that it would not be due to a small chance sampling and that it would not be reversed the next time. Again, suppose a study has been carried on with correlations and it is found that the correlation between the amount of discussion in a group and the gain, other things being equal, is 0.62 whereas the correlation between the amount of lecture and gain, other things being constant, is 0.30. The problem is "Would this difference always be found or is it due only to the chance sampling?" Suppose that this were based upon a study of 100 cases, then using the formula given on page 240 it is apparent that the probable error of a correlation of 0.62 when based upon 100 measures is 0.04, and the probable error of a correlation of 0.30 based upon 100 measures is 0.06, using this time the formula for the probable error of the difference (which is exactly like that for the standard deviation) we have

$$\begin{aligned} \text{P.E.Difference} &= \sqrt{0.04^2 + 0.06^2} \\ &= \sqrt{0.002 + 0.004} \\ &= \sqrt{0.006} \\ &= 0.08 \end{aligned}$$

If a difference is to be surely significant, it should be 4.4 times the probable error, that is, it should cover a range in this case of 0.35. The actual difference was only 0.32 or four times the probable error of the difference. The chances are, however, 142 to 1 with a measure which is four times its probable error. This is very good, but it is not large enough for practical certainty. The study of this example ought to be carried further, tried out on a larger sampling of persons so as to reduce the

probable error of correlations, and to make any difference which does appear practically certain.

It is very easy to be misled when dealing with differences unless their probable errors or standard deviations are calculated. An official in a religious organization was comparing the work of a man in one state with the work of a man in a different state. He found that on some items such, for example, as the amount of time spent in raising money, there was a large difference between the subjects, and he assumed these were, therefore, significant. Upon other items, such, for example, as the amount of time spent for study, he found that the difference was very small and he regarded this as insignificant. The error in his procedure became clear when the standard deviations for each difference had been found. The variation among the men within each state in the amount of time spent in raising money was so great and the standard deviation so large that this apparently large difference amounted to only two times its own standard deviation and was, therefore, not surely significant. The apparently small difference, however, was based upon figures very homogeneous with the two states, therefore possessing a very small standard deviation. The difference, small in itself, turned out to be six times its standard deviation and, therefore, surely significant.

d. Limitations of Reliability

While the importance of the calculation of measures of unreliability should not be minimized, there are certain misinterpretations which should be prevented, if possible. The theory of reliability is based upon the assumption of normal distribution. If the experiment deals with traits or quantities which almost surely do not fall into a normal distribution, or if the sampling be so small, say less than 25, that it bears little, if any, resemblance to the normal curve, then measures of reliability are rather meaningless. Moreover, it is impossible by the calculation of the limits of practical certainty to make allowances for errors which exist in the bias of the investigator or in such a choice of subjects that they cease to be a fair sample of the group for whom he would like to make the results applicable.

EXERCISES

1. Find each of the measures of central tendency and of variability for the series of test scores obtained by application of a test to this class.

2. Procure an unsigned statement of the approximate number of times a church meeting of some sort has been attended within the past two months, from a large number of persons. Find the mean and S. D. Graph the results.

3. Give two tests to some group and find the correlation both by the rank method and the direct method. Interpret the results in three different ways.

4. Given the following data, find whether ministers' salaries are more closely related to pastoral ability or to executive ability, when length of service in the ministry is held constant.[1]

Salary with executive ability	$r = 0.93$
Salary with pastoral ability	$r = 0.60$
Length of service with executive ability	$r = -0.07$
Length of service with pastoral ability	$r = 0.12$
Salary with length of service	$r = 0.89$
Executive ability with pastoral ability	$r = 0.57$

Interpret this in terms of the per cent of factors leading a minister to get an increase in salary which are included in ratings on his executive ability but not included in ratings on his pastoral ability.

5. Find the actual self-correlation of the test constructed in Exercise 5 for Chapter VI.

6. Find the correlation between the results of that test and an intelligence test given to the class, and then correct the result for attenuation to see how far the tests, if perfectly reliable, would still measure the same thing.

7. *a.* Knowing length of service and a pastor's ratings in executive ability and in pastoral ability, how accurately could you predict his salary? What weight would be given to each factor? (Use data in question 4.)

b. Assuming average salary $2,100, with an S. D. of $750
average length of service 8 years S. D. 2.5
average pastoral ability rating 3 S. D. 1.0
average executive ability rating 3 S. D. .9

What is the most probable salary for a man who has served 5 years, is rated 5 in executive ability, and 6 in pastoral ability? What is the range of practical certainty on this prediction?

c. Could you be practically certain that a man who has served 10 years, is rated 3 in executive ability and 6 in personal ability, would get more or less?

8. Carry through the necessary computations on your own problem. Be sure you state the reliability of your results.

[1] Data taken from MOXCEY, "Success in the Christian Ministry," Teachers College Bureau of Publications, 1922.

CHAPTER VII

PRESENTATION OF THE RESULTS OF EXPERIMENTATION

The discovery of facts never in itself brings progress. It is only when somebody makes use of the facts which have been discovered that progress results. James Harvey Robinson, in "The Humanizing of Knowledge," well emphasizes the obligation of the scientist to take the rest of the world into consideration. The gap between what is known and what is practiced is one of the most serious handicaps which present-day civilization faces.

An experimenter in religious education at the present time is likely to be confronted with this sort of situation. His results may be either what is expected by everybody in advance of his experiment, or they may be a challenge to the ordinary assumptions. If the results are what everyone expected, it may be argued that he might as well not have done his experiment. If the results are unexpected, it is quite probable that no one will believe them anyhow. If this situation continues, no one will be more to blame than the experimenters themselves. It is not enough to find truth. Somehow that truth must be made to function in the experience of the individuals for whom it is found.

1. Choice of Audience

One of the first decisions which the experimenter should make concerns the need that he expects to see met by his experiment. It will frequently help him to think of it in terms of the persons whom he expects to help. At this there may be an indignant protest from some who are interested in "Truth for its own sake." This sounds like a noble ideal, but it hardly bears careful investigation. Countless absurd trivialities might consume the energies of men, were there any value in purposeless fact finding.

Some of these tidbits of information can be found in the com-

pilations of 10,000 facts which few people care to know. Every serious-minded scientist has some purpose for his truth finding. It is true that that purpose may not be one of an immediate and practical sort. It may fulfill a need which he and some of his peers recognize as important while no one else apparently sees the significance in the sort of thing he is doing. The plea, however, that the finder of truth need have no purpose in mind save the discovery of fact is hardly in keeping with the situation presented by the many urgent needs of modern civilization and the few people who are able to help find the sort of truth which will lead toward richer life.

Among the persons of whom the experimenter may think as being persons who will most probably use the results of his experiment may be suggested the following:

1. The experiment may be intended to fulfill a scientific need, to advance knowledge in a sphere in which other people are laboring. The results then should be prepared for the use of trained experimenters, of persons who will be able to build the next step upon previous advances.

2. Results may be designed for the use of technically trained experts in religious education. These may be curriculum makers or board administrators, or teachers in college and graduate schools who face the task of formulating the best theory and practice in moral and religious education.

3. The results may be designed to appeal directly to the average church-school worker who takes his task seriously and wishes to do better next week than he did during the week previous.

4. An experiment may be designed for the use of everyone interested in education in a popular sense. That is, it may be designed to help parents and community leaders and persons who are concerned with education in a non-professional way.

5. The experiment may be selected to help the persons who are going to be educated. Probably too few experiments have been performed with the object of helping children to educate themselves. The pupils have been assumed as raw material rather than as persons cooperating in a process. Perhaps it will

seem desirable to perform experiments directly with the purpose of helping children to participate more efficiently and with a larger amount of satisfaction to themselves.

If the persons who are most likely to make use of the results have been determined, the next question is the formation of the results into a shape which will serve this group. In all probability the decision as to the persons to be served and the form in which the results are to be stated should be made before the experiment is begun. It will influence the kind of information collected and the processes by which the information is treated as well as the form in which the report is made.

2. Written Reports

The first and most frequent type of report is that which is a written dissertation, thesis, monograph, or similar study. In preparing such a written report several suggestions may be of value:[1]

1. The problem upon which the study bears should be stated clearly and explicitly.

2. The sources of data should be summarized.

3. The method should be described so carefully that any other scientist can repeat the experiment. This means that the people involved must be studied and reported with much more precision than has been usual. Experiments carried on with certain groups or by experimenters of unusual personality are surely subject to influence which might make repetition in exactly the same form very difficult. The tests and other materials used should be included in the report or at least samples of them should be given.

4. Each question or subquestion stated in the definition of problem should receive a separate discussion somewhere, and the best answer to it which the study has to contribute should stand out. It should not be necessary to hunt all through a report, gathering up pieces of information here and there which have some bearing on the question the report is supposed to answer.

[1] Valuable suggestions are offered in Monroe and Johnson, "Reporting Educational Research," University of Illinois *Bulletin* No. 25, Urbana, Ill., May 18, 1925. Price 50 cts.

5. If the discussion is long or involved, it is desirable to summarize the conclusions.

6. If there are any findings which are not in agreement with popular expectation or which would have an unusually wide bearing upon practice, this should be pointed out and special attention should be given to the justification of such a conclusion, and the changes which it would require in present practice. It cannot be assumed that if a truth is stated its application will be clear to the reader.

7. Technical terms should be used in the sense in which they are ordinarily understood. If new terms are coined or phrases are used in a special meaning within the study, attention should be called to these modifications.

8. Details of form should be punctiliously observed. A standard method of stating footnotes and bibliographical references should be accepted and carefully followed. The captions of tables should be given at the top, those of graphs should be given at the bottom. Tables, figures, and examples should be so labeled that they can be understood without reference to the accompanying text. The text, on the other hand, should be so complete that it can be read by persons who do not care to turn aside for the tables and graphs.

9. The dictates of English composition should not be ignored. Unity, coherence, clearness, force, interest, and even beauty should be considered. There is nothing about scientific writing which makes it impossible, indiscreet, or irreverent to say things in attractive literary style.

3. Use for Publicity

The results of the experiment may be presented not in a formal written report but in a form more suitable to publicity purposes. It may involve newspaper writing, an exhibit, or public meeting in which the results are to be discussed. One of the most effective methods for presentation in such a situation is the use of posters and charts. Statistical results are usually much more clearly interpreted and much more likely to be studied by the general public if they are presented in the form of graphs rather than in the form of tables. Sometimes the sector graph is used,

in which a circle is divided into proportionate amounts. This is very frequently done when it illustrates the way in which money is spent. Table XXV shows a graph which illustrates how the board whose administration problem was discussed on pages 224 to 237 should divide its funds. Sometimes a bar diagram is used. This is illustrated in Table X. The graphic frequency distributions illustrated in Table VII may be regarded as another type for presentation. Striking points on any graph may be indicated by heavy arrows pointed toward them.

The following recommendations with reference to graphs, taken from the report of a committee on standards for graphic representation, may be helpful:[1]

1. The horizontal scale for curves should usually read from left to right and the vertical scale from bottom to top.

2. Where possible, represent quantities by linear magnitudes as areas or volumes are more likely to be misinterpreted.

3. For a curve the vertical scale, whenever practicable, should be so selected that the zero line will appear on the diagram.

4. If the zero line of the vertical scale will not normally appear on the curve diagram, the zero line should be shown by the use of a horizontal break in the diagram.

5. The zero lines of the scales for a curve should be sharply distinguished from the other background guide lines, especially when the zero point is not at the bottom or extreme left.

6. It is advisable not to show any more background lines than necessary to guide the eye in reading the diagram.

7. The curve lines of a diagram should be sharply distinguished from the ruling.

8. In curves representing a series of observations, it is advisable, whenever possible, to indicate clearly on the diagram all the points representing the separate observations.

9. Figures for the scales of a diagram should be placed at the left and at the bottom or along the respective axes.

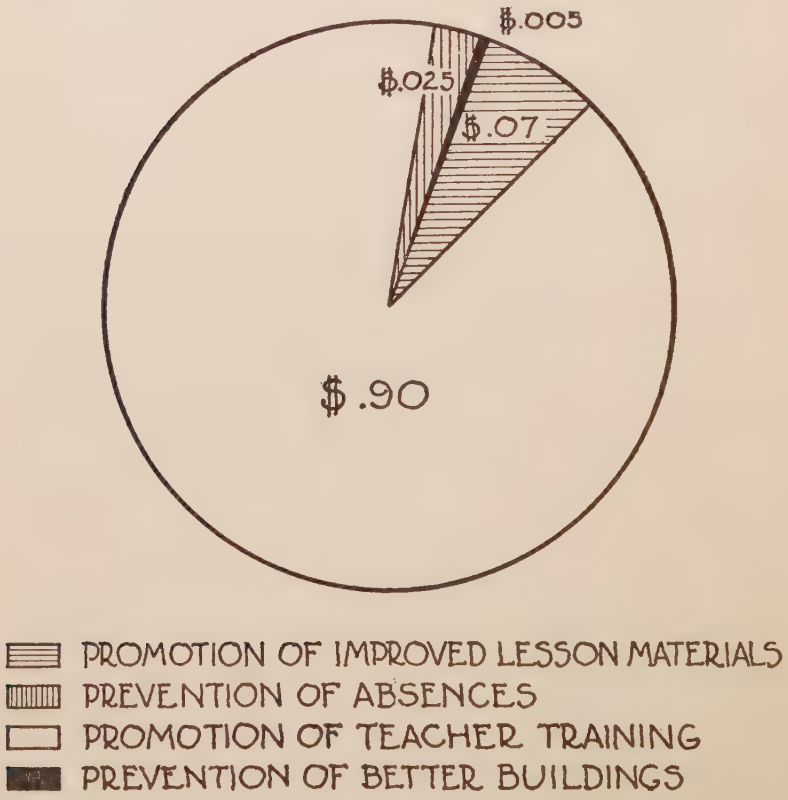
10. The title of a diagram should be made as clear and

[1] For further material on graphs see BRINTON, "Graphic Methods for Presenting Facts," The Engineering Magazine Co., New York, 1917.

complete as possible. Subtitles or descriptions should be added if necessary to insure clearness.

11. When several items are being compared the item of chief interest may be made more striking than the others.

TABLE XXV. — HOW THE BOARD OF RELIGIOUS EDUCATION SHOULD SPEND EACH DOLLAR



12. In graphing two or more bars or curves for comparison make their zero lines coincide.

13. Do not use a percentage curve when it is wished to show the actual amounts of increase or decrease and do not use an amount curve when it is wished to show the per cents of increase or decrease.

When results are presented not in the form of simple graphs but in the form of posters or cartoons for popular appeal, a careful study should be made of the principles of appeal, form, and beauty, which govern the laying out of proportions, choice of color schemes, and other relevant factors. The fact that a series of posters presents the results of an experiment does not justify some of the travesties on art which it is not difficult to discover. A study of the best modern advertising practice affords excellent guidance in poster construction for publicity purposes.

Moving pictures are not infrequently used to record just what has taken place at significant moments in an experiment. The recent development of inexpensive reels which can be operated by the amateur opens up fields for the recording of experimental results and the presentation of those results to an interested audience, which may well be utilized.

4. Education of Those Who Participate

One of the best "reports" may come through participation in the experiment itself. Frequently, the men who get most out of an experiment are those who have done it. Many community surveys have been made in which the presentation of results might have been entirely omitted and still the survey would have been of tremendous value. People who work at it learn things as they work, which influence their thinking and practice. Even at the cost of a certain amount of efficiency in obtaining scientific results, it is usually wise to bring into the experiment as many as possible of the people who are likely to use the results. This involves taking them in from the very beginning. They have a contribution to make in formulating the problem which is to be studied. It sometimes appears that a problem formulated in order to be of service to a certain group of people proved not to be of service because some element was left out of the formulation which they regarded as particularly important. So far as possible, the constituency should participate in choosing what tests should be given. It is discouraging, to say the least, to come to the end of a long experiment only to find the group who might have been significantly helped by the experiment saying, "But you did not test this particular thing!"

A report was made to a group of Y M C A secretaries recently. On the basis of this preliminary report an experiment for the next year was outlined. One of the persons upon whom the success of the experiment particularly depended said, "I will be glad to cooperate if I can be 'in' on the interpretation of the results. I do not want to gather up material in an experiment and then have someone else come back and tell me what to do." As statistical processes grow more complicated and the techniques of experimentation become more refined, it must undoubtedly seem to lay observers that the scientist in educational procedure takes his plates into the darkroom and comes out with a picture which did not seem to be at all evident on the face of the material which he took into the darkroom. The lay observers do not trust the picture. Something may have happened in all of these technical processes which has distorted the meaning.

This does not mean that technical processes are not to be used. It does mean that as rapidly as possible, people are to be encouraged to understand and appreciate the task of the technician in experimental work. Perhaps there will always be certain people who are primarily policy makers and others who are primarily executants, but in so far as people who are actually doing the work come to feel that they have been not only concerned with mechanical detail but have had a real part in formulating policies and finding factors, the ideal of democracy is more nearly approached.

5. Twenty-five Pitfalls

The following list of intellectual immoralities will be of interest to every experimenter regardless of his purpose or technique. These have been formulated by the Character Education Institution,[1] and point out with keen insight and utter frankness some of the points at which human frailty tends to obstruct the progress which those who have devoted themselves to the search for important truth hope to achieve. They are suggested at this point, with the hope that experimentation and measurement in religious education will so profit by the experience of others, that these suggestions will never be needed.

[1] Chevy Chase, Washington, D. C. Milton Fairchild, Director.

Intellectual Immoralities

a. Carelessness in observations, "sloppy work."
b. Inaccuracy in determining units to be counted in statistical research.
c. Slovenliness in logic, fantastic explanations.
d. Generalizing beyond one's data.
e. Confusing opinions with knowledge.
f. Confidence in the results of research in disregard of weakness in proof and verification.
g. Contentment with "discussion."
h. Poor judgment in research plan and procedure.
i. Wavering interest, flitting attention, attracted by peculiar superficialities.
j. Egoism allowed to crowd one to the invention of "new" theories for personal distinction.
k. Inventing interesting theories for the sake of using them in conversation or selling them in books, articles, and lectures.
l. Pride allowed to result in persistent belief in a theory for which one has been given credit.
m. Formulating a hypothesis on weak bases of facts, and then becoming blind to facts in opposition.
n. Emotionalism during research, "I believe" instead of "I have proved."
o. Adjusting theories to popular likes and dislikes.
p. Opposition to proof of another's theories because of jealousy.
q. Opposition to a theory merely because of ignorance and stupidity, "I can not see how."
r. Rushing into print with a report of research work that justifies no conclusions.
s. Degenerating into a propagandist of an unproved hypothesis, instead of being true to the research purpose of discovering the truth.
t. Cowardice in supporting a verified generalization because it is unpopular and conflicts with selfish interests.
u. Impatience, unwillingness to proceed step by step through a research.
v. Indulgence in dense verbiage for the sake of appearing superlearned.
w. Ignorance of the mechanism of instruments of precision, which results in their use when out of order.
x. Popularizing tentative generalizations for the sake of personal publicity.
y. Resort to the authorities, or to sarcasm and ridicule, against data, arguments, and verifications.

6. Religion and Scientific Research

In such warnings and reproaches there is implied a sense of some higher values. The scientist imposes restraint upon himself, not to glory in his bonds, but because he thereby serves what is, to him, a nobler end. He labors in the faith that he is doing his part in humanity's endeavor to answer two of life's greatest questions: "What is this universe in which we find ourselves?" and, "How shall we live for the things of most worth?" Many creeds have brought many answers. Science is not a new orthodoxy with a new set of practices and answers. It is, in contrast, religious. That is, it is Spirit and Truth. It is Worship. It pervades the soul of man until he cannot walk blindly or carelessly. It sets him in a world of strangeness, and bids him wonder that anything can happen. The sophistication of science makes the commonplace incomprehensible. The pursuit of any truth leads relentlessly, as Carlyle has suggested, to "that vast valley of humiliation where the wisest of men must feel themselves forever children, gathering pebbles on the shore of an endless sea."

Devotion to scientific research means more than a solitary quest for Truth. It demands in rich measure that each shall labor for all. Each scientist must work with tools which he did not build, at tasks others have well begun. He must deny himself easy results and surround every conclusion with caution, not because this is in itself pleasant, but because he works for unknown others. "So others shall take patience, labor, to their heart and hand, from his hand, and his heart, and his brave cheer; God's grace fructify through him, to all." It was upon what happens today among men of science that one of the noblest pictures of the Kingdom of God was painted, Royce's "Beloved Community."

EXERCISES

1. Prepare the results of the experiment you have carried on during the course in satisfactory form for publication.

2. Select three Ph.D. dissertations and choose the main material from each which could be presented through the public press. Write the "story."

3. Prepare a graph to illustrate the situation on this country as described in the answer to the first question submitted in the article "An Evaluation of Current Religious Education," appearing in *Religious Education*, February, 1925, Vol. XX, p. 56.

4. Contrast the attitude of the scientific mind watching a baby raise its hand to its mouth, with that of the unsophisticated common-sense individual. What are the gains and losses of the scientific wonder?

APPENDIX

Suggested Problems for Experimental Investigation — The Construction of a Sample Scale — Statistical Tables — Glossary

SUGGESTED PROBLEMS FOR EXPERIMENTAL INVESTIGATION

The following questions are by no means an exhaustive list. Most readers will quickly think of a score of others more important than some included in the list. The concluding forty were formulated by the Y M C A. The aim has been to stimulate thought with reference to possible experiments. Some of these suggested are suitable for an average Sunday-school teacher, others ought to discourage the trained research worker. Some have been stated in limited form, applied to specific age groups and situations, others have been left in general form, needing later definition and limitation. It is suggested that attempts to outline, perhaps in terms of the methods presented in Chapter II, just how each of these might be investigated will prove fruitful in the development of ability to see the research possibilities in new situations.

1. Will teachers endeavoring to use a project viewpoint produce pupils who, after 3 years of experience, show more desirable growth than is shown by pupils taught by more traditional recitation methods?

2. Will children 8 years of age of average intelligence, be more apt to want to dramatize a story which has been read to them, or an identical story which has been told to them?

3. Do institutions of religion show conservatism on modern social questions in proportion to the wealth they possess?

4. Are people who are well acquainted with the Bible more or less apt to be prejudiced on religious questions than are people with less biblical training?

5. Are people who hold rather consistently radical religious and social views very largely people who were brought up under certain kinds of home training?

6. What sort of hymns (*e.g.*, slow, lively, familiar, new, modern in

vocabulary, major, minor, jazzed, etc.) appeal to thoughtful children in an ordinary church school, as most valuable?

7. What type of church advertising (*e.g.*, colored posters, church bulletins, hand bills, car cards, newspaper notices, form letters, personal calls, some combination of these) will produce the best results in terms of number of people attracted in proportion to expense in time and money?

8. What is the effect of strong emotional states on ability to understand, to remember, to size up a situation, to think, and to act skilfully?

9. Is a church best served, in the long run, by excellent preachers who do little pastoral work, by inferior preachers who do excellent pastoral work, or by men who do fairly well at each?

10. What other symptoms of emotional instability are likely to be associated with persons who hold very strong prejudices upon moral questions?

11. Does the telling of Old Testament stories to junior age boys and girls make more or less difficult the development of an idea of God like that set forth in the New Testament?

12. In the average church situation today would better results in attendance, interest, and Christian conduct be obtained by substituting group discussion for the Sunday sermon?

13. Is it true that racial and international attitudes are likely to be the result of occasional striking experiences, usually in childhood?

14. Is more effective change in attitude brought about through groups which meet primarily for the discussion of religious and ethical problems, or through the discussion of religious and ethical aspects of questions in groups gathered together primarily for other purposes?

15. Does missionary giving on the part of children tend to increase or to decrease their enthusiasm for missions after adolescence?

16. Are there significant differences in glandular make-up or nervous stability among persons who are religious radicals and those of equal intelligence, in somewhat similar environments, who are fundamentalists?

17. Does learning to play basketball in a spirit of fairness, under the supervision of a Christian leader, have a significant effect upon the attitudes of fairness which boys show in other life relationships?

18. Is more remembered from a lecture given to a group of college students or from a discussion conducted upon the same problem, with a similar group, for the same length of time?

19. Is it better to expose a child deliberately and gradually to the seamy side of life, or to allow him to grow up somewhat sheltered from poverty, sickness, crime, vice, etc.?

20. Does giving a child playthings which are regarded as its very own tend to increase respect for property of others, or to create envy and selfishness?

21. Within which of the large denominations are the leaders most clearly agreed upon important social issues, and what is the nature of that agreement or difference?

22. Is native intelligence a more important factor in determining what is learned in Sunday school than is teaching method?

23. What tests, if given to high-school students, will with most accuracy select those capable of the best leadership of groups of younger boys and girls?

24. Who makes the moral ideals of children? What is the relative influence of parents, playmates, school teachers, Sunday-school teachers, editors of papers taken in the home?

25. Are the people who are members of religious organizations different in physique, intelligence, personality, emotionality, education, etc., from people who are not members of these organizations?

26. Are there significant differences in physique, intelligence, personality characteristics, emotionality, temperament, etc., between the members of one religious denomination and those affiliated with other denominations?

27. Does a given type of sex education make any difference in the thought and practice of children and adolescents?

28. Given equal presentation, are children 10 years of age more likely to respond generously to appeals for foreign need, or to appeals for local relief?

29. Are children trained in Chinese mission schools less patriotic than those trained in Chinese government schools?

30. Do 10 hours of Bible class work add more to the ability of children 12 to 14 years of age in making ethical discriminations than would 10 hours spent on hikes with an equally able leader?

31. Will paid teachers in the City Sunday school accomplish better results in terms of pupil-information, pupil-attitudes, and pupil-conduct than would be accomplished by volunteer teachers under similar circumstances?

32. Does one's idea of God have a direct connection with the ideal one held of mother or father, when a child?

33. Is the financial support of missionary work increased, diminished, or unaffected by missionary addresses revealing the unchristian impact of the West upon the East, as regards diplomacy, commercialism, race prejudice, etc.?

34. Is better religious growth made in classes which are alike in age, school grade, Bible knowledge, native intelligence, interests, and purposes? Does homogeneity increase or diminish valuable learning about social relationships?

35. Do moving pictures impress the consequences of certain ways of living upon the minds of children more effectively than would stories or sermons?

36. Does knowledge of the common standard of morality increase among Boy Scouts in the community more rapidly than among other boys of equal initial ability and equally favorable home and school environment?

37. How is delinquency related to such possible causes as low intelligence, school maladjustment, broken homes, sickness in childhood, type of neighborhood, number of available playgrounds, economic status, church-school attendance, education, play groups?

38. Do meetings of church officials held in private homes tend to be freer from uncordial relationships than do similar meetings held in church rooms?

39. Is more or less remembered from a lesson taught in a group containing both boys and girls about 15 years of age, than would be remembered if the groups were taught similar material in separate groups?

40. Does excellent physical equipment tend to attract a different group from that attracted by a similar program in less adequate building and surroundings?

41. Is there a significant difference in the tendency to action about an undesirable situation, when the appeal is made with a stirring address, and when the situation is faced through discussion?

42. Do pupils in classes which are entirely self-governed do more effective study than is done by pupils in classes receiving regular and definite assignments worked out by the teacher?

43. Will an attempt to use the project approach and to utilize situations as they arise, lead to the ignoring of important information, attitudes, and conduct habits, which would be included by a more systematic teacher of equal ability?

44. In the present state of religious education, are more effective results secured by the average teacher when given (*a*) no helps at all, (*b*) a variety of suggestions from which choice must be made, (*c*) a general outline with encouragement to modify this, or (*d*) specific instructions embodying the best present knowledge to be followed as carefully as possible?

45. Does compulsory chapel contribute to or detract from the religious interest of college students?

46. Are the attitudes of children most likely to be influenced by stories that deal with modern characters or ancient ones, stories about adults or about children, continued stories or short stories, stories read or told, stories of good acts, bad acts, or combinations, stories that are fanciful or close to reality?

47. Do cigarettes produce any clear effect upon mental ability or moral judgment in persons over 15 years of age?

48. What is the effect of Boy Scout training upon attitudes toward war?

49. What size of unit, in terms of area and population, is the most effective for the supervision of religious education?

50. Does petting, either monogamous or promiscuous, tend to decrease the probability of a happy and successful marriage?

51. At what mental age do children gain a clear conception of the symbolism involved in a typical Christmas or Easter pageant?

52. Do people who have been converted in a definite experience tend to possess different attitudes or stronger loyalty to religious points of view, than do persons who have grown up in similar environments but without such experiences?

53. Does thrift, taught in a club, carry over into the lives of boys and girls over a long period of time?

54. What is the effect upon health, industrial or mental efficiency, and enjoyment, of spending Sunday at the church, or spending Sunday on a hike in the woods, or a combination, as a general summer practice for young people 16 to 30 years of age?

55. Do pupils of given age groups, after a single session of experience, tend to choose teachers more wisely than would adults after one long interview with the prospective teachers or leaders?

56. Is the success of a given club leader predictable, at the beginning of his services, on the basis of (*a*) previous experience, (*b*) intelligence, (*c*) punctuality, (*d*) ability to outline a good club-meeting program, (*e*) emotional history records, (*f*) knowledge of teaching principles, (*g*) ability to suggest topics which would interest the group, (*h*) general impression made upon the supervisor, or (*i*) some combination of these measures?

57. Which of the following, or other techniques for opening a class in religous education with pupils of high-school age will, on the whole, lead to the selection of the most important and worth-while questions for study: (*a*) informal conversation, (*b*) discussion of what the group has done in previous years, (*c*) use of true-false statements to start discussion; (*d*) placing books, pictures, curious objects, etc., around the room, and allowing time for examination of and informal conversation about these; (*e*) actual trial of several types of curricula during an experimental period; (*f*) questions as to what the group would like to do; (*g*) choice by the pupils among three or four suggested plans outlined by the teacher; (*h*) preliminary study by the entire group of what other classes have done; (*i*) taking of a survey test and noting deficiencies?

58. Does baptism by immersion tend to make a person's loyalty to the church more permanent than do other forms of baptism?

59. Do children compelled to attend church tend to enjoy it, after some years, more than do children of equal home status and intelligence, who attend only when they choose?

60. What are the significant differences in personality make-up (*e.g.*, health, glandular development, intelligence, temperament); in church experiences (a love attraction within the group, election to positions of pres-

tige, opportunities offered, courses studied, teachers with whom associated); and in outside experiences (schooling, type of occupation, sources of recreation) between those adolescents who remain active in church and those who lose interest in any given organization and community?

61. Is it true that theology is formed largely by hymns?

62. Does a church-school orchestra increase or decrease the sense of reverence among the pupils?

63. What is the effect of a missionary study class upon gifts for benevolence, attendance at other missionary lectures, and upon world-mindedness?

64. Is attitude toward any important issue of Christian ethics in young people 12 to 17 years of age most influenced by straight facts, by vivid stories, or by some combination?

65. Do responsive readings contribute ideas or feelings of worship to adults or children?

66. What is the effect of a community week-day school of religion upon Sunday-school attendance?

67. Over a period of 5 years, do boys 11 to 16 years of age gain more in biblical knowledge, ethical discrimination, desirable attitudes upon personal and social questions, interest in service enterprises, and conduct as observed by parents and teachers, from a program which has set before them certain standards and a well-built program, or from one which has encouraged them to be completely self-determining?

68. Which are the most effective of several possible methods of dealing with a bully?

69. What is the effect of praying for the sick, if the invalids do not know about the prayer?

70. Assuming that home influences are constant, is there evidence that pupils with Sunday-school training develop into better characters than do others of equal intelligence and general socio-economic status?

71. Is there an optimum amount of responsibility to give to children which will tend to produce independence, but not to overburden them with responsibilities which are too serious for them to face?

72. Are marked differences between children in willingness to conform to social needs set by the time the children are old enough to come to school, so that only rarely thereafter can they be changed?

73. Do pupils develop, so far as their judgment and that of those competent to observe them is concerned, more desirable religious attitudes in worship based upon modern poetry, aesthetic dances, rituals of other religions, and similar features, than they do in traditional services?

74. How do the results in ethical information, judgment, and actual conduct when children are given 3 hours a week for 6 months a closely correlated program of Sunday school, week-day school, and club work, compare with the results obtained by similar children, given the same

amount of time with equally capable leadership, when the three programs are wholly independent of one another?

75. What are the comparative results of several common methods for training volunteer leaders, such as: community training schools, occasional courses based upon ordinary texts in child study and teaching method, leaders' discussion conferences, or personal supervision and conference?

76. Can religious education be carried on as effectively (using religious in the sense of those concerns about God and man which are not paralleled by denominational distinctions) in public schools as in Sunday schools?

77. What aspects of religion (theology, worship practices, acquaintance with the personalities of religious leaders, service enterprises in religious organizations) tend to have the largest effect upon everyday conduct of pupils 9 to 12 years of age?

78. Are young men better served by educational, physical, social, and religious activities correlated within one group, or by separate enterprises to which they can affiliate themselves as they wish?

79. Is there a difference between the ability of persons who believe they can do all things through Christ, to do the things they really want to do, and the ability of persons who believe that what they can do is up to themselves and who rely only on themselves, to do similar things?

80. Do young people who read the Bible daily lose their tempers less often during the day than do people who do not have such a devotional practice?

81. What would be the relative effect, in determining life motives as evidenced by occupations, of a course in Bible study given to college students, or a course in modern social problems, or a combination course?

82. Does taking the part of a certain kind of character in a dramatic production tend to influence the normal behavior of the amateur actor in other groups?

83. Would better results, in terms of the religious and ethical and general personality development of children, be secured by 2 years of effort devoted wholly to parent education, all present week-day and Sunday-school projects being abandoned, than would be secured by the attempt to develop character in these educational groups for children themselves?

84. Is worship best suited for meeting the needs of children 9 to 12 years of age, when conducted by children or by adults, when planned by children or by adults, when largely formal and ritualistic or when largely informal?

85. Will a man who honestly tries to follow Christ in the world today be prospered in material things?

86. What sort of subjects would eighth-grade pupils in a church school choose for their curriculum, if they had had the experience of making choices and abiding by the consequences, since they entered the school?

87. Are marriages between people whose ability, interests, and charac-

teristics are alike likely to be more successful than are marriages between people whose characteristics differ?

88. What proportion of Christian life-work decisions made during childhood or early adolescence tend to cause misery, to be of negligible influence, to lead definitely toward happiness and satisfaction?

89. Is it more desirable for a class to carry forward one benevolent enterprise over a period of years, or to give a little help to each of a number of varied enterprises each year?

90. What are the most effective methods of dealing with an individual who is possessed by a strong prejudice which interferes with the group progress?

91. If pupils studied no Bible at all until they were 17 to 20 years of age, and then were given a year of thoroughgoing study of biblical history, literature, and interpretation, would they come out with a better understanding of the Bible than is possessed by pupils of equal ability who study a little bit here and a little bit there?

92. Which of the experiments in cooperation between religious and social agencies seem to have produced the best results in terms of the present activities within the community?

93. What is the relation between religious activities in college and previous religious experiences?

94. What differences are there between the fair-mindedness and the professional success of ministers whose training has included Greek and those who have not had such training, those who have had systematic theology and those who have not, those who have studied parish problems and those who have not, those who have studied much biblical material and those who have studied little, those who have studied religious education to a large extent, and those who have had little or no such training?

95. What relation is there between the stabilizing of a country church and the efficiency of agriculture?

96. What criteria, if reported in a yearbook, are within the range of possibility for the average religious organization, and would give the most significant information about the real religious and moral contributions of the organization?

97. What are the relative effects of two systems of character training, alike in general, except for the fact that one includes prizes, emblems, honors, awards, and recognitions, while the other takes account only of the intrinsic worth of an activity to the pupil?

98. Does the constant stirring up of individuals about problems (*e.g.*, international adjustments) with which he has few if any effective contacts, and about which he can only form a strong conviction, lead most frequently toward (*a*) greater interest in such problems and ability to contribute to them when the opportunity comes, (*b*) a general blunting of interest in a

sense of hopelessness, (*c*) sentimentality which finds no useful outlet, (*d*) some other result?

99. Would a given religious organization make more progress in its service over a long period of time, if it reduced its overhead supervisors to a few competent individuals who would give full time to the direction of research and experimentation, educating the constituency in the process?

100. What will be the relative value, in terms of all measurable indices of growth, of the study of such curricula as are suggested by each of the following classes?

(*a*) This class visits a movie each week. The vivid, gripping portrayals of virtue and vice are used instead of story material for Sunday discussions.

(*b*) The textbook for this class is the Sunday paper. In it they find evidence of the work of God in the world today. National problems, economic injustices, world religions, poverty, crime, benevolence, even the progress of literature and science are to be found in its voluminous reading matter.

(*c*) This class uses a weekly magazine, the *Literary Digest*, *New Republic*, *Christian Century*, or some similar publication for its quarterly. It makes its business the endeavor to formulate a Christian attitude on the problems and progress of the modern world.

(*d*) The leader of this class brought to the first session an armful of all the magazines he could find at the newsstand and purposely came to the class late so that the pupils would have an opportunity to start reading along the lines of their choice. Out of this grew a study which lasted several months and included an evaluation of the real worth, from a Christian point of view, in various types of reading matter.

(*e*) This class recognized that the Hebrew myths were less attractive and less stimulating to children of modern America than are the myths of the American Indians, the old Norse heroes, the Greeks, or the Romans. They chose to study the meaning of life, the story of creation, character of God, value of prayer and worship in these other attempts of mankind to find God.

(*f*) This class recognized that the progress of science was not only appealing to modern youth but represented some of the most important religious developments of the past century. The characters they studied were, therefore, men of science.

(*g*) This class, in response to a need of some nursery children for a playhouse, spent their class time in the cooperative enterprise of making a playhouse to be given to the small children.

(*h*) This class selected four important national problems and spent one-fourth of a year in the discussion of each. They chose problems of

race conflict, problems of economic adjustment, problems of sex relations, and the prohibition problem.

(*i*) This class made its main choice for the year the study of other religions and the comparison of these with Christianity.

(*j*) This class obtained slides of the great paintings in Christian art (it might have been better could they have gone to the museum of art in their city). They found that to understand each picture led them into much study of church history, Christian symbolism, and the lives of the great Christian painters. In addition to the information they found that the course opened up gateways of appreciation which were really worshipful.

(*k*) This class was interested in reading and writing poems expressive of the religious experience of the present age. They found that high-school pupils expressed more vital and intimate aspirations of religious life in an attempt to write religious poems than in any other way.

(*l*) This class recognized the dearth of hymns which were suitable for modern youth. They went at their problem in two ways. They sought to find or construct fitting music for some of the poems which they liked best and also to write new words for some of the hymns, the music of which they loved.

(*n*) This class in a large city spent every other week in visiting some other church and church school. They tried to prepare themselves to enter sympathetically into the background of other religious groups. This involved study of the history, creeds, and institutions. They tried to find in each the things which made it valuable for its adherents.

(*o*) This class made its project personal and community health. This involved eventually a study of civic relations, the responsibility of one for another, and so on. Consideration of disease led naturally into discussions of poverty, alcoholism, and so on.

(*p*) This class centered all its attention on writing and producing plays and pageants. In addition to the conduct training which came through cooperation, there were values in the information they obtained in making scenery, costumes, background, and so forth, and also in the choice of plays as true to life and of religious significance.

(*q*) This class took seriously its function of interpreting life from the Christian point of view. Its raw material was the public school curriculum. The teacher visited science classes and then helped the pupils later to integrate their science and their religion. The teacher visited English classes and later helped the pupils appraise the characters portrayed. History classes needed to see the religious purpose in its relation to history. Classes in civics and economics needed to conceive a Christian social order.

(*r*) This class tried to help its members find the best life work. Voca-

tions, especially the newer service vocations, were studied with reference to the qualifications they demand and the opportunities they offer.

(*s*) This class tried to help its members find out how to have a good time in their community. What movies should one see? What books and magazines should one read? What people should one go with? How intimately?

(*t*) This class rendered its service to the school by editing a mimeographed school magazine. They tried to be of service to every other class in the school. Not only the traditional news and notices were included but if classes were studying India, some of the members looked up material on India and wrote an article which would help in the class study. They investigated suggested philanthropic enterprises and tried to present the comparative merits of several to which classes might give their money. They became the servants of all in the search of helpful material.

(*u*) This class visited institutions within their city. They visited toy shops, factories, transportation systems, the homes of rich and poor, hospitals, and churches. They tried to decide which things they were glad to have in their city and which they would like to change.

101. What are the most significant indices for prediction of the life of a given suburban community 10 years hence?

102. What is the genetic psychology of a certain emotional drive? For example, the appearance of anger and fighting in babies has been fairly well analyzed by Watson and Jones. By what steps and stages does this develop into the anger reactions which one finds at the ages 5, 10, 15, and 50?

103. How far are preferences, interests, attitudes, or ideals influenced by stories (told or read), newspapers, magazines, and novels?

104. Should a supervisor work primarily on the basis of giving specific concrete suggestions by which workers may solve present problems, or on the basis of asking questions, leading the local workers to analyze their problems and to discover resources for themselves?

105. How do the attitudes toward persons of other races develop and change? Is the process different for children from that which goes on with adults? What is the effectiveness of negative adaptation, direct conditioning, verbal appeal, social imitation, etc.?

106. What consequences of observing motion pictures of certain typical kinds can be observed in children of various age levels? Is there a tendency for an exciting picture to leave individuals wrought up and incapable of effective intellectual or skilled work?

107. What tests can be developed to measure given aspects of progress in religious and moral growth? For example, what measures can be developed to predict sensitiveness toward the happiness and comfort of other people, especially those at remote social distance?

108. What is the relationship between cost of necessary building equipment and upkeep and the character contribution from such activities as bowling, billiards, handball, lobby-checkers, educational courses, etc.?

109. What differences appear in school dishonesty among boys who are of equivalent intelligence and home background, some of whom have no Y M C A training, others 1 year, 2 years, 3 years, and so forth?

110. In what ways are members of Y M C A's atypical? How do its members and leaders compare with the rest of the community in intelligence, physical development, economic status, religious background, social prestige, community influence, and so on?

111. What ought to be the relationship of the Y M C A to delinquent boys? What have been the experiences of courts which have paroled boys in care of Y M C A secretaries? In what ways could this service be improved?

112. What factors contribute to or interfere with successful marriage adjustments? This might involve a study of successful and unsuccessful cases. A special problem is raised by "heavy necking" in relation to normal sex adjustment.

113. Is it true that fundamental emotional patterns change little after the first ten years of life?

114. In organizations dependent upon voluntary contributions, how far is the attitude toward local industrial situations determined by sources of financial support?

115. Precisely what activities are involved in the work of a Boys' Work Secretary (or Religious Work Secretary, Physical Director, Educational Secretary, etc.)? This investigation would employ the recognized job analysis technique and might have to be confined to associations of a given type or communities of a given type.

116. What form of reporting the experiences that a leader has with a group will lead to compilations which will give the largest amount of guidance in future situations? Can a technique as definite as the social case workers' report be developed?

117. What appear to be the most significant causes of misconduct? This would involve the study of three carefully equated groups: (*a*) misconduct, (*b*) average, and (*c*) unusually excellent. It is particularly necessary that the theories apparent in existing case studies of delinquent and mentally or emotionally abnormal children be checked by control groups.

118. At what points do the attitudes of the average citizen of the United States have any direct bearing upon international relationships? Probably such investigation should be based upon analysis of a great many actual cases. Behind it lies this problem: Are most of the efforts to interest people in international problems futile so far as any hope of real expression

on the part of the average citizen is concerned? Are international affairs problems for a few experts?

119. What types of prayer are common among young men 18 to 25? What is the relationship between these prayer-types and the patterns of emotional history? What relationship exists between these prayer-types and present conduct?

120. Should the Y M C A adult education activities seek to be accredited by existing academic agencies? What is the influence of accrediting upon service rendered?

121. What is the function of increased information and insight upon the conduct and interests of young men? Does character demand any emotional or volitional elements in addition to the intellectual ones? Of course it is understood that increased information and insight into consequences bring inevitable emotional accompaniments. The question is, is that adequate or must there be the addition of extrinsic dynamic factors?

122. Through what process of self-study may a community be led in order to bring about the best understanding of conflicting interests and the most satisfying integration and change?

123. What type of records if kept by local associations over many years would provide the best basis for scientific knowledge about character formation?

124. What features of summer camps for boys are most closely connected with desirable results? On this problem many significant data have already been collected and await tabulation and statistical study.

125. What are the best techniques for interviewing? Are there elements which are common to personnel interviews, religious interviews, charting, vocational counsel, group inquiries, psychoanalysis, the confessional, and social case work?

126. What should be included in an index of resources for leaders of boys' clubs? How can such a wealth of material be classified and indexed so it can be of most service to groups building their own programs? Is classification on the basis of age levels necessary? Does size and type of community require reclassification? Should the institutional channel be emphasized, grouping together activities connected with home, with school, with church, and vocation?

127. What elements in intelligence, health, emotional history, and educational training are most closely related to success in Y M C A leadership? This field may have to be subdivided: *e.g.*, Do successful physical directors know more about anatomy, physiology, etc., than do unsuccessful ones from the standpoint of their total product?

128 What is the significance of motives in health education? What incentives lead not only to desirable health habits but also to desirable character growth?

129. Exactly what are the limits of transfer of a moral principle? This is perhaps the most important problem in program making which the Association faces. It requires research on carefully controlled groups of children who are given training in situations *a*, *b*, *c*, and *d* and tested in other situations. The factors which determine the amount of spread are almost unknown except under such ambiguous terms as "identical elements" and "generalization."

130. What has happened in present attempts to secure an attitude of greater mutuality, to get rid of the notion of the superiority of America in relation to mission enterprises? What has been and will be the result on interest and on financial support?

131. How can the eventual success of leaders be predicted? Can we recognize and select in advance only those possessing certain desirable qualities of leadership?

132. Under what conditions have adults with deep-dyed prejudices, racial, religious, economic, educational, organizational, made radical changes in attitude? Why? How?

133. How do adult committees function? Where do ideas originate? How are elements which might interfere with committee success handled?

134. What has actually happened to *all* of the men and boys who 10 years ago participated in certain activities? For example: a summer camp period; a deeply stirring religious conference; gym activities; educational classes. What does each person now think of the value of those experiences?

135. In interracial conflicts where and how have adjustments been made so that the vague fears of everyone concerned were dispelled?

136. How are young men 18 to 25 actually spending their time? What needs and interests does this reveal?

137. How do systematic and other types of Bible study compare with certain other activities in their contribution to character (*e.g.*, study of a newspaper)?

138. What is actually being done by men who are now rendering the most service to industrial communities? How are they affecting human waste, strikes, attitudes of conflict? Is it possible to conserve the highest values in a standardized factory?

139. Should the Association offer to young men many special interest groups, or should it be expected that each group will develop a variety of activities? "Omnibus Club?"

140. What are the essential differences and likenesses of character built up under fear trends or drives and those built up under love trends or drives?

THE CONSTRUCTION OF A SAMPLE SCALE

The method of arriving at the quantitative rating of specimens is based on the Cattell-Fullerton theorem that equal differences are equally often noticed.[1] This has been proven true, within limits, for irregular geometric areas. If 60 per cent could see that *B* was larger than *A*, but 70 per cent could see that *C* was larger than *A*, then one could be reasonably certain that *C* was larger than *B*. When the agreement is less than 60 per cent or more than 90 per cent, the theorem becomes somewhat less reliable, but between those limits, differences do tend to be proportionate to the number of judges who can notice them. The proportion is not direct, but is based on the normal distribution curve. If 50 per cent think action I better than action II and 50 per cent would disagree, the two actions are probably of equal merit. If 75 per cent think that I is better than II, then act I lies further toward the high end of the scale, a distance along the base

line of the curve equal to 0.68 standard deviations. If 80 per cent think I better than II, then the distance along the base of the curve is 0.85 S. D.'s. So each per cent of area may be interpreted in terms of distance along the base line, that is, real difference in merit.

The procedure in the construction of such a sample scale might well be as follows.

a. Obtain 100 specimens of the act, trait, quality, or complex, which is to be rated.

b. Have these ranked by at least 12 competent judges.

c. Average the ranks, re-rank on the basis of averages, and pick out for the final scale 10 instances, including one of the lowest, one of the highest, and the remainder at fairly equal steps between. Discard instances on which there was confusion or wide variability of rank.

d. Submit these 10 samples in chance arrangement, for ranking to several hundred judges, representative of the group later to use the scale.

e. Fill in the following table, giving at each point the per cent of the judges believing the instance given at the head of the column superior to the one mentioned at the left. Let I represent the best and X the poorest.

[1] Thurston, in "Equally Often Noticed Differences," *Jour. Educ. Psy.*, XVIII, p. 289, May, 1927, points out that this is true only when samples are alike in range or spread of excellence as measured by many judges.

TABLE XXVI. — FORM FOR RECORDING PER CENT OF JUDGES PREFERRING ONE SPECIMEN TO ANOTHER, FOR 10 SAMPLES

	Per cent rating I higher than	Per cent rating II higher than	Per cent rating III higher than	Per cent rating IV higher than	Per cent rating V higher than	Per cent rating VI higher than	Per cent rating VII higher than	Per cent rating VIII higher than	Per cent rating IX higher than
II.	——								
III.	——	——							
IV.	——	——	——						
V.	——	——	——	——					
VI.	——	——	——	——	——				
VII.	——	——	——	——	——	——			
VIII.	——	——	——	——	——	——	——		
IX.	——	——	——	——	——	——	——	——	
X.	——	——	——	——	——	——	——	——	——

f. By means of the following table, find the distance from one sample to another in terms of S. D. units. Compare each with the one directly above it in value. Thus, if 78 per cent think IX better than X, the distance between them is 0.77 S. D.'s. If 92 per cent think VIII better than IX, that distance is 1.41 S. D.'s, etc.

g. If it is desired to get the most accurate value possible, the distance between samples I and III, II and IV, III and V, etc., may be found and checked against the previous values. So likewise where steps of three, I to IV, II to V, III to VI, etc., do not involve more than 95 per cent agreement, they may be used. In general, the larger the step, the less weight should be given to the scale distances determined by using it. For most purposes the simple first-step comparisons under *f* will be adequate.

h. Assign an arbitrary value to the poorest instance of the scale. Probably it would be possible to conceive of instances which would rate lower than the worst one in this scale of 10. Hence the lowest should be given some value above zero. Five is ordinarily a good value. Then the specimen IX will be a value above 5, found by adding to 5 the S. D. distance between IX and X. VIII will be still higher, and its value is found by adding to the value just found for IX the S. D. distance between VIII and IX. So the process may be continued up the scale.

TABLE XXVII. — CONVERSION OF PER CENT OF JUDGES BELIEVING A GIVEN SAMPLE SUPERIOR TO A SECOND SAMPLE, INTO THE DISTANCE IN S. D. UNITS BETWEEN FIRST AND SECOND SAMPLES

Judges believing one better, per cent	S. D. distance	Judges believing one better, per cent	S. D. distance
50	0.00	76	0.71
51	0.02	77	0.74
52	0.05	78	0.77
53	0.08	79	0.81
54	0.10	80	0.85
55	0.13	81	0.88
56	0.15	82	0.92
57	0.18	83	0.95
58	0.20	84	0.99
59	0.23	85	1.04
60	0.25	86	1.08
61	0.28	87	1.12
62	0.31	88	1.18
63	0.33	89	1.23
64	0.36	90	1.28
65	0.39	91	1.34
66	0.41	92	1.41
67	0.44	93	1.48
68	0.47	94	1.56
69	0.50	95	1.65
70	0.53	96	1.75
71	0.55	97	1.88
72	0.58	98	2.05
73	0.61	99	2.33
74	0.64	99.5	2.58
75	0.68	99.9	2.90

i. Print the scale, placing the best instance first, then the next best, next best, etc., and print in front of each its value as determined in the preceding steps. The result is a scale which any judges can use, and by means of which coarse qualitative judgments can be brought to a more reliable quantitative basis. Any new instance to be judged need only be written on a card and moved up and down on the scale until a point is found at which all those above it seem better, all those below it poorer. The numerical value of this step is a fair method of representing the value of the act being judged.

TABLE XXVIII. — SQUARES AND SQUARE ROOTS TO 100

Number	Square	Square root	Number	Square	Square root	Number	Square	Square root
1	1	1.000	36	12 96	6.000	71	50 41	8.426
2	4	1.414	37	13 69	6.083	72	51 84	8.485
3	9	1.732	38	14 44	6.164	73	53 29	8.544
4	16	2.000	39	15 21	6.245	74	54 76	8.602
5	25	2.236	40	16 00	6.325	75	56 25	8.660
6	36	2.449	41	16 81	6.403	76	57 76	8.718
7	49	2.646	42	17 64	6.481	77	59 29	8.775
8	64	2.828	43	18 49	6.557	78	60 84	8.832
9	81	3.000	44	19 36	6.633	79	62 41	8.888
10	1 00	3.162	45	20 25	6.708	80	64 00	8.944
11	1 21	3.317	46	21 16	6.782	81	65 61	9.000
12	1 44	3.464	47	22 09	6.856	82	67 24	9.055
13	1 69	3.606	48	23 04	6.928	83	68 89	9.110
14	1 96	3.742	49	24 01	7.000	84	70 56	9.165
15	2 25	3.873	50	25 00	7.071	85	72 25	9.220
16	2 56	4.000	51	26 01	7.141	86	73 96	9.274
17	2 89	4.123	52	27 04	7.211	87	75 69	9.327
18	3 24	4.243	53	28 09	7.280	88	77 44	9.381
19	3 61	4.359	54	29 16	7.348	89	79 21	9.434
20	4 00	4.472	55	30 25	7.416	90	81 00	9.487
21	4 41	4.583	56	31 36	7.483	91	82 81	9.539
22	4 84	4.690	57	32 49	7.550	92	84 64	9.592
23	5 29	4.796	58	33 64	7.616	93	86 49	9.644
24	5 76	4.899	59	34 81	7.681	94	88 36	9.695
25	6 25	5.000	60	36 00	7.746	95	90 25	9.747
26	6 76	5.099	61	37 21	7.810	96	92 16	9.798
27	7 29	5.196	62	38 44	7.874	97	94 09	9.849
28	7 84	5.292	63	39 69	7.937	98	96 04	9.899
29	8 41	5.385	64	40 96	8.000	99	98 01	9.950
30	9 00	5.477	65	42 25	8.062	100	1 00 00	10.000
31	9 61	5.568	66	43 56	8.124	101	1 02 01	10.050
32	10 24	5.637	67	44 89	8.185	102	1 04 04	10.100
33	10 89	5.745	68	46 24	8.246	103	1 06 09	10.149
34	11 56	5.831	69	47 61	8.307	104	1 08 16	10.198
35	12 25	5.916	70	49 00	8.367	105	1 10 25	10.247

TABLE XXIX. — VALUE OF r CORRESPONDING TO EACH VALUE OF ρ (ACCORDING TO THE FORMULA $r = 2 \sin (30^\circ \times \rho)$ IN WHICH $\rho = 1 - \frac{6\,\Sigma D^2}{n(n^2-1)}$)

ρ	r	ρ	r	ρ	r	ρ	r
0.01	0.010	0.26	0.271	0.51	0.528	0.76	0.775
0.02	0.021	0.27	0.282	0.52	0.538	0.77	0.785
0.03	0.031	0.28	0.292	0.53	0.548	0.78	0.794
0.04	0.042	0.29	0.303	0.54	0.558	0.79	0.804
0.05	0.052	0.30	0.313	0.55	0.568	0.80	0.813
0.06	0.063	0.31	0.323	0.56	0.578	0.81	0.823
0.07	0.073	0.32	0.334	0.57	0.588	0.82	0.833
0.08	0.084	0.33	0.344	0.58	0.598	0.83	0.842
0.09	0.094	0.34	0.354	0.59	0.608	0.84	0.852
0.10	0.105	0.35	0.364	0.60	0.618	0.85	0.861
0.11	0.115	0.36	0.375	0.61	0.628	0.86	0.870
0.12	0.126	0.37	0.385	0.62	0.638	0.87	0.880
0.13	0.136	0.38	0.395	0.63	0.648	0.88	0.889
0.14	0.146	0.39	0.406	0.64	0.658	0.89	0.899
0.15	0.157	0.40	0.416	0.65	0.668	0.90	0.908
0.16	0.167	0.41	0.426	0.66	0.677	0.91	0.917
0.17	0.178	0.42	0.436	0.67	0.687	0.92	0.927
0.18	0.188	0.43	0.446	0.68	0.697	0.93	0.936
0.19	0.199	0.44	0.457	0.69	0.707	0.94	0.945
0.20	0.209	0.45	0.467	0.70	0.717	0.95	0.954
0.21	0.219	0.46	0.477	0.71	0.726	0.96	0.964
0.22	0.230	0.47	0.487	0.72	0.736	0.97	0.973
0.23	0.240	0.48	0.497	0.73	0.746	0.98	0.982
0.24	0.251	0.49	0.507	0.74	0.756	0.99	0.991
0.25	0.261	0.50	0.518	0.75	0.765	1.00	1.000

TABLE XXX. — PREDICTIVE INDICES FOR CERTAIN CORRELATION COEFFICIENTS

$$\text{P. I.} = 1 - \sqrt{1 - r^2}$$

Correlation	P. I.	Correlation	P. I.	Correlation	P. I.
0.00	0.00				
0.01	0.00	0.36	0.06	0.71	0.30
0.02	0.00	0.37	0.07	0.72	0.31
0.03	0.00	0.38	0.08	0.73	0.32
0.04	0.00	0.39	0.08	0.74	0.33
0.05	0.00	0.40	0.08	0.75	0.34
0.06	0.00	0.41	0.09	0.76	0.35
0.07	0.00	0.42	0.09	0.77	0.36
0.08	0.00	0.43	0.10	0.78	0.37
0.09	0.00	0.44	0.10	0.79	0.39
0.10	0.01	0.45	0.11	0.80	0.40
0.11	0.01	0.46	0.11	0.81	0.41
0.12	0.01	0.47	0.12	0.82	0.43
0.13	0.01	0.48	0.12	0.83	0.44
0.14	0.01	0.49	0.13	0.84	0.46
0.15	0.01	0.50	0.13	0.85	0.47
0.16	0.01	0.51	0.14	0.86	0.50
0.17	0.01	0.52	0.15	0.87	0.51
0.18	0.02	0.53	0.15	0.88	0.53
0.19	0.02	0.54	0.16	0.89	0.54
0.20	0.02	0.55	0.16	0.90	0.56
0.21	0.02	0.56	0.17	0.91	0.59
0.22	0.02	0.57	0.18	0.92	0.61
0.23	0.03	0.58	0.19	0.93	0.63
0.24	0.03	0.59	0.19	0.94	0.66
0.25	0.03	0.60	0.20	0.95	0.69
0.26	0.03	0.61	0.21	0.96	0.72
0.27	0.04	0.62	0.22	0.97	0.76
0.28	0.04	0.63	0.22	0.98	0.80
0.29	0.04	0.64	0.23	0.99	0.86
0.30	0.05	0.65	0.24	1.00	1.00
0.31	0.05	0.66	0.25		
0.32	0.05	0.67	0.26		
0.33	0.06	0.68	0.27		
0.34	0.06	0.69	0.28		
0.35	0.06	0.70	0.29		

TABLE XXXI. — INTERPRETATION OF CORRELATION IN TERMS OF DISPLACEMENT [1]

Showing, for various amounts of correlation, the chances in 100 that the second measure of an individual will be in the same "tenth" of a distribution, not displaced more than 1 "tenth," 2 "tenths," etc. The sixth line is read as follows: In the case of a .50 correlation the chances are only 26 in 100 that an individual's second score will be in the same "tenth" of the distribution as his first score; the chances are 69 in 100 that his second score will not be displaced by more than 1 "tenth"; etc.

Coefficient of correlation	Number of "tenths" displacement							
	0	1	2	3	4	5	6	7
0.00	19	53	77	91	97	99.2	99.8	99.9 +
0.10	20	55	79	92	98	99.4	99.9	
0.20	21	58	82	94	98	99.6	99.9 +	
0.30	22	61	85	95	99.0	99.8	99.9 +	
0.40	24	64	88	97	99.5	99.9 +		
0.50	26	69	91	98	99.7	99.9 +		
0.60	29	73	94	99.2	99.9 +			
0.70	34	81	97	99.8	99.9 +			
0.80	41	89	99.2	99.9 +				
0.90	55	98	99.9 +					
0.95	71	99.9						
0.98	91							
1.00	100							

[1] From OTIS, "Statistical Method in Educational Measurement," The World Book Co., Yonkers, 1925.

TABLE XXXII. — PER CENT OF INDEPENDENT CAUSAL FACTORS COMMON TO TWO MEASURES YIELDING CERTAIN CORRELATIONS [1]

Correlation	Causal elements common, per cent	Correlation	Causal elements common, per cent	Correlation	Causal elements common, per cent
0.00	0	0.36	28	0.72	51
0.01	1	0.37	28	0.73	52
0.02	2	0.38	29	0.74	52
0.03	3	0.39	30	0.75	53
0.04	4	0.40	30	0.76	54
0.05	5	0.41	31	0.77	55
0.06	6	0.42	32	0.78	55
0.07	7	0.43	32	0.79	56
0.08	7	0.44	33	0.80	57
0.09	8	0.45	34	0.81	58
0.10	9	0.46	34	0.82	59
0.11	10	0.47	35	0.83	60
0.12	11	0.48	35	0.84	61
0.13	12	0.49	36	0.85	61
0.14	12	0.50	37	0.86	63
0.15	13	0.51	37	0.87	64
0.16	14	0.52	38	0.88	65
0.17	15	0.53	38	0.89	66
0.18	15	0.54	39	0.90	67
0.19	16	0.55	40	0.91	69
0.20	17	0.56	40	0.92	70
0.21	18	0.57	41	0.93	72
0.22	18	0.58	42	0.94	73
0.23	19	0.59	42	0.95	75
0.24	20	0.60	43	0.96	77
0.25	20	0.61	44	0.97	80
0.26	21	0.62	44	0.98	83
0.27	22	0.63	45	0.99	88
0.28	23	0.64	45	1.00	100
0.29	23	0.65	46		
0.30	24	0.66	47		
0.31	25	0.67	48		
0.32	25	0.68	48		
0.33	26	0.69	49		
0.34	27	0.70	50		
0.35	27	0.71	50		

[1] Taken from NYGAARD, "A Percentage Equivalent for the Coefficient of Correlation," *Jour. Educ. Psy.*, February, 1926.

TABLE XXXIII. — PER CENT OF TOTAL NUMBER OF CASES (AREA OF NORMAL CURVE) FALLING ABOVE GIVEN SCORES IN S. D. UNITS

S. D. Units are $\frac{\text{Standard Deviation}}{10}$ with zero at −5 S. D.

Score in S. D. units	Per cent above this score	Score in S. D. units	Per cent above this score
0	99.999971	51	46.02
1	99.999952	52	42.07
2	99.999992	53	38.21
3	99.99987	54	34.46
4	99.99979	55	30.85
5	99.99966	56	27.43
6	99.99946	57	24.20
7	99.99915	58	21.19
8	99.9987	59	18.41
9	99.9979	60	15.87
10	99.9968	61	13.57
11	99.9952	62	11.51
12	99.9928	63	9.68
13	99.989	64	8.08
14	99.984	65	6.68
15	99.977	66	5.48
16	99.966	67	4.46
17	99.952	68	3.59
18	99.931	69	2.87
19	99.903	70	2.28
20	99.865	71	1.79
21	99.81	72	1.39
22	99.74	73	1.07
23	99.65	74	0.82
24	99.53	75	0.62
25	99.38	76	0.47
26	99.18	77	0.35
27	98.93	78	0.26
28	98.61	79	0.19
29	98.21	80	0.13
30	97.72	81	0.097
31	97.13	82	0.069
32	96.41	83	0.048
33	95.54	84	0.034
34	94.52	85	0.023
35	93.32	86	0.016
36	91.92	87	0.011
37	90.32	88	0.007
38	88.49	89	0.0048
39	86.43	90	0.0032
40	84.13	91	0.0021
41	81.59	92	0.0013
42	78.81	93	0.0009
43	75.80	94	0.0005
44	72.57	95	0.00034
45	69.15	96	0.00021
46	65.54	97	0.00013
47	61.79	98	0.00008
48	57.93	99	0.000048
49	53.98	100	0.000029
50	50.00		

GLOSSARY

Terms often used in experimentation and measurement are here explained in the sense in which they are used in this book. So far as agreement among experimenters permits, all terms have been used in their customary significance. For more complete discussion of certain terms, consult index.

Accomplishment Quotient; A. Q.: The ratio obtained by dividing the educational quotient (E. Q.) by the intelligence quotient (I. Q.). The A. Q. indicates the extent to which a pupil's learning is keeping pace with his ability. If it is 100 he is doing as much as should be expected of an average person of his intelligence. If it is less than 100 he is accomplishing too little, if it is over 100 he is doing better than do average persons with his capacity.

Achievement Test: A measure of what has been learned, not a measure of capacity but of progress actually made, results achieved.

Algebraic Addition or Subtraction: The signs of the numbers are taken into account. The sum of the negative numbers is subtracted from the sum of the positive numbers in algebraic addition.

Alienation, Coefficient of, K: A number which indicates how far a correlation is from perfect prediction. If the alienation coefficient were 1.00, then the prediction of scores in one series from knowledge of scores in the other would be sheer guess. If it were zero, the prediction would be absolutely perfect.

Analytical Score: A score which breaks up a total into several parts. For example, a gross intelligence score may be broken up into factors of speed, power, ability to handle abstract problems, etc. Score for Bible knowledge may be broken up into Old Testament score and New Testament score.

Approaching a Limit: Coming closer and closer and closer but never actually reaching the limit point.

Arbitrary Value: A value which depends on the judgment of a person or a small group, which cannot be proven to be correct by any mathematical process but in the last analysis has to rest back on somebody's say-so.

Assumption: An idea, statement, or relationship which is taken for granted either because it seems obvious or for the sake of developing a process to fit certain cases. A statistical process is only as true as its weakest assumption.

Attenuation: Literally, "sag," or "droop." It is used to show what happens to a correlation if that correlation is based upon unreliable measures. The correction for attenuation shows what the correlation would probably have been if the measures had been reliable. Because of attenuation,

correlations are too low. Correction for attenuation tends to raise them but frequently raises them too much.

Average Deviation (A. D.): A measure of the differences between members of a group which has been tested. It is the average distance of each score in the group from the middle. If groups are very homogeneous the average deviation is small; if the groups show much deviation and spread then the average deviation is large.

Base Line: The horizontal line in a graph, from which vertical distance is measured. In a normal curve it is customary to mark off units of the trait which is being measured along the base line.

Battery of Tests: A collection of several tests used in combination.

Causal Factor: One of the things which helped to cause a given result. There are usually many causal factors which work together in bringing about every situation.

Central Tendency: A measure of the middle or center of the group. This is the best single indication of the group's position.

Chance: Influenced by a large number of unknown causal factors, each having a relatively small part.

Check: To go over work again in order to see whether the same answer is found, thus making sure that the first work was correct.

Coefficient: A number which has a certain statistical meaning. An index. The coefficient of correlation measures relationship; the coefficient of alienation measures the amount of chance in a prediction, etc.

Comprehensive Test: A test which covers in complete, or at least in fairly representative fashion, what the test is supposed to cover. A test which puts too much emphasis upon some unimportant points and omits or gives too little weight to some important things is not comprehensive.

Completion Questions: Statements in which some important words or phrases are omitted and which test the student's knowledge by his attempt to fill in these blanks.

Composite Score: A score which is made up by combining in some fashion the results of several tests.

Concomitant Learnings: Learnings which were not directly planned for or expected, but which appear along with the learnings directly sought.

Constant Error: In most experiments the many slight errors balance one another fairly well. If one operation runs a fraction of a second too long the first time, it may be a fraction too short the second time. The people who are feeling unusually depressed are balanced by people who are feeling unusually cheerful. If, however, there is some factor that affects the outcome of the experiment which is not balanced and neutralized in this way, it becomes a constant error. It must be measured, where possible, and allowance made for it.

Control Group: A group used in an experiment to make sure that the re-

sult was due to the particular operation being studied. The control group should be like the experimental group in every way. It should have everything done to it which is done to the experimental group, except for the one operation. The result should appear in one group and not in the other.

Controlled Answer: That type of test or questionnaire which suggests possible answers to be checked or otherwise indicated. The person taking it may not write out any answer he pleases, but must use one of the suggested responses.

Controlled Experiment: An experiment which takes such careful account of everything which might possibly influence the outcome, that the one cause can be undoubtedly connected with the one outcome. Sometimes it is done through control groups, sometimes statistically. It is usually the sort of thing done in careful scientific laboratory research, rather than that done by the field experimenter.

Correlation, Coefficient of, r: A measure showing the relationship between two sets of measures both of which have been obtained on the same group. There may be, for example, a tendency for persons or groups making high scores in one measure to make similarly high scores in the other, and for persons low in the first trait to make similarly low scores in the other. If such a parallelism is perfect, the correlation is 1.00. Any correlation above 0.80 is considered high. If the relationship is pure chance, and there is no connection between standing high in one trait and standing high or low in the other, the correlation is zero. Any correlation between − 0.40 and 0.40 is considered low, and little better than chance. If people or objects standing high in one measure stand low in the other the correlation is negative. If this is a perfect inverse relationship, the correlation is − 1.00.

Criteria, Criterion: The standard by which a test is judged. It should be self-evident that the criterion is an index of the thing to be measured.

Criterion Score: The score showing where an individual or group stands, with reference to the criterion. This is used as a basis for figuring out how the test itself should be scored, so as to yield results as much as possible like the criterion.

Critical Score: A figure which separates a group so that those above it are distinctly different from those below it. The boiling point of a substance is a critical point for it. The passing mark in a college is a critical score.

Cumulative Effect: The result obtained when one operation builds upon the preceding one, so that each new one starts, not from the ground floor, but from all that has happened previously.

Data (plural of *Datum*): Figures, descriptions, similar evidence upon which experimental conclusions are to be based.

Decile: The points which divide a distribution into tenths of its total number are decile points. The decile may be thought of as the distance in score between one of these points and another. Thus, a person falling in the third decile has at least 20 per cent of the group poorer than he is, and at least 70 per cent who are better.

Delinquent: A term applied to persons under 21 years of age who have committed offenses, usually minor ones, which necessitate court action.

Diagnostic: Tending to show the causes for the result. A diagnostic test tells not only the strong points and the weak points, but indicates the sort of difficulty underlying this condition.

Displacement: The amount of movement in score or position brought about by retaking a test, or the difference between true position and one predicted. If displacement is small, everyone appears in the second case just about where he did in the first one. If there are large chance elements, displacement is great.

Distribution Curve: A graph which shows how many persons or groups or objects fell at each step of the measure. Usually the scores are measured along the base line, and the number of persons indicated vertically.

Duplicate Equivalent Test Forms: Two or more forms of a test which will, if given to the same group under the same circumstances, yield approximately the same results. Usually the forms are different, but each question in one is paired by a question in the other which requires the same sort of ability to answer it. Duplicate equivalent tests are very valuable when it is desired to test before and after applying some operation.

Educational Quotient (*E. Q.*): A ratio obtained by dividing educational age (E. A.) by chronological age. The educational age is the age of the average person in schools tested, who knows as much about subjects included in tests, as does this person. It is sometimes found by averaging subject-ages. Thus, a pupil equal to the average 10-year-old in arithmetic, the average 14-year-old in reading, 12-year-old in geography, 11-year-old in penmanship, 13-year-old in spelling, etc., might have an average educational age of 12 years. If he were 10 years old chronologically, his E. Q. would be $12/10 = 1.20$. See Intelligence Quotient and Accomplishment Quotient.

Element: Each question, or item, or subdivision of question or item in a test is called a test element. It is the smallest test unit.

Empirical Formula: A formula worked out to fit what experience shows usually happens, not based on strict mathematical analysis.

Equated Groups, Equivalent Groups: Groups set up for the purpose of an experiment, and so chosen as to be alike in all the matters which might influence the outcome of the experiment. If exactly the same thing were done to two or more equated groups the outcome should be exactly the same.

Equation: An expression involving numbers or symbols, or both, which shows two ways of saying the same thing. $2 + 2 = 4$ is a simple equation.

Error: Sometimes error is used in its ordinary sense of mistake, omission or slip. It has another meaning in scientific work, which is more nearly related to impurity in a chemical. It is a causal element operating where it is not wanted. The errors in an experiment confuse it, so that it is difficult to connect the one cause with the one result. The errors in a test influence the score, although they are not the sort of thing the test is supposed to be measuring.

Essay Examination: The old type of test question, in which the pupil writes out in his own words and way the answer to questions requiring discussion.

Extreme Scores: Scores which are at the very ends of the distribution, far from the average. Particularly extreme scores are those which are separated from the rest of the group, which lie in a different range entirely.

Fake: To pretend, to give the appearance of what is not so. A test which is non-fakable is one which makes it impossible for a person to show himself up as good or bad, according to his desire.

Final Test: The test given after an operation. The one given before is often called the initial test.

First-order Partials: Partial correlation coefficients in which one element has been held constant. These tell the relationship between one trait and another, which would exist if all the subjects tested were alike in some third trait.

Formula: A sentence explaining a relationship, expressed in symbols (mathematical shorthand) instead of in words.

Free Association: The expression of a person's reverie, letting ideas drift along, come as they will, connect up however strangely they may. This is contrasted with controlled association in which a person is directed to let his mind run in certain channels, to think of colors, or friends, or names of countries.

Free Response, Natural Response: This is used in contrast to controlled response. A free or natural response is unlimited. A person is placed in the situation and allowed to react as he chooses.

Frequency Distribution: A statistical table indicating how many persons, groups, or objects fell within each step of the measure. It answers the question, "How frequently did each score (or range of scores) appear?"

Gaussian Curve: The normal frequency curve.

Graphic Distribution: See *Distribution Curve.*

Group: Used to indicate the single person or several persons or large number of persons upon whom some experiment is tried, or to whom tests and measures are applied.

Halo Effect: The tendency of a rater to be influenced by his general liking for, or dislike of, a person he is supposed to rate. Such a liking makes all ratings high, a dislike pulls them undeservedly low. The feelings toward the person as a whole is carried over into specific traits which are really quite unrelated.

Hold Constant: When a particular trait is held constant it means that every one in the group is alike in that particular trait.

Homogeneity: Used in its ordinary sense, meaning all-of-a-kind-ness. Used as the opposite of variability or spread, to mean that the group were much alike, and all made scores close to the average.

Independent: Unrelated, having no causal factors in common. When one goes up or down or disappears entirely, the others are not in the least influenced.

Index, Indices: A sign, indication, or system, which signifies a meaning beyond its intrinsic content.

Initial Test, I. T.: The test given at the beginning of the experiment, as contrasted with the final test at the end.

Integer: A whole number, with no fractions or decimals attached.

Intelligence: Variously defined by many writers. Usually indicates inborn capacity to make adjustments, do abstract thinking, size up situations quickly, etc. It is not an acquired result of education, it is inborn brightness. May be defined as the sort of thing measured by the best intelligence tests.

Intelligence Quotient, I. Q.: The ratio obtained by dividing mental age by chronological age. Mental age (see below) is obtained from tests.

$$\frac{\text{M. A.}}{\text{C. A.}} = \text{I. Q.}$$

The intelligence quotient stays relatively constant from year to year; it shows not absolute ability, but ability relatively to age. It is customary in figuring I. Q. to use 16 as the maximum C. A., even for adults.

Interchangeable Units: Steps on a test scale which are exactly alike, so that a gain of one point high on the scale means exactly the same as a gain of one point in the middle or at the lower end. Grams are interchangeable. One may be taken out and another added, and the scales still balance. Arithmetic problems are not so interchangeable. To miss one may be much less serious than to miss another.

Intermittent Application: Used to describe an experimental method in which a certain influence is tried for a while; then stopped, then later tried again, then followed by a rest, then tried again, etc.

Isolate: To free from errors, or other impurities. A causal factor is isolated when its influence is clearly separated from every other influence.

K. Coefficient of Alienation: see *Alienation.*

Linearity: Applied to a graph to mean tendency to fall along a line.

Line Graph: A graph in which the length of a line indicates the comparative strength or achievement, or gain, etc. Sometimes also called a bar graph.

Matching and Pairing Questions: New type questions which ask the subject to indicate which of the words in one list belong with each of the words, phrases, or illustrations in another list.

Mean, Average (Mn): The common measure of the middle of a group, found by adding the scores together and dividing by the number in the group.

Median (Med): The score attained by the middle individual of the group, or, the score which has just as many below it as above it.

Mental Age (M. A.): A measure of intelligence is often translated into mental age, meaning, as good as the average person of that age. It should be carefully distinguished from the intelligence quotient, or I. Q. Mental age increases from year to year, at least up to adult life. I. Q. stays fairly constant. Mental age tells how able the individual is, I. Q. tells whether his brightness is keeping up with his age.

Mode: The most frequent score.

Multiple Choice Questions: New type questions which suggest several options and measure the ability of the subject by the kind of answer he chooses.

New Type Questions, or Examinations: Controlled answer questions which confine the responses of the subject to certain limits. These are in contrast to essay-type questions in which one may write what one pleases. New type test questions include true-false, multiple-choice, completion, matching, and ranking questions.

Normal Curve, Normal Distribution, Normal Frequency Curve: A graph showing how many times each score or result occurred, when based on an infinite number of cases, influenced by chance factors. See page 183.

Norms: Standards based on previous use of the test, which show what may be expected of certain well known or typical groups. Often norms are based on average children of each age or grade.

Objective, Objectivity: Independent of the personal or subjective element. Giving the same result regardless of who may use it.

Operation, Operating Factor: Here used as whatever is done in an experiment. It is whatever happens between the "before and after taking." It may be consciously planned by the experimenter, or certain operating factors may come in without his consent.

Optimum: The most desirable, the best. The optimum weight for each of several factors to be used in predicting something else, is the weight which will make the prediction most perfect.

Order Distribution: An arrangement of scores with the highest at the top, the lowest at the bottom, and the rest in their proper positions between.

Pair: Individuals or questions are paired, when each is matched as carefully and completely as possible by some other.

Partial Correlation: A measure of the relationship between two traits or measures, when one or more other measures are held constant. In first-order partial equations (see above) only one additional measure is held constant. In second-order partials, the relationship is studied when two measures have been equated for the individuals of the group. In third-order partials, the relationship between two traits may be studied with the groups assumed to be alike in three factors. These things which are held constant, are sometimes said to be "partialled out." They no longer influence the relationship. It is as though every one in each group were exactly alike with reference to the traits partialled out.

Partial Regression: Just as it is possible to predict scores from simple correlations (zero order), so it is possible to predict most probable scores from knowledge of several measures, and their partial correlations. Partial regression makes use of partial correlations in telling what the score in one trait is likely to be, from knowledge of scores in other traits. Partial regression coefficients are the weights assigned to each of these traits which are used in the prediction.

Partial Standard Deviation: A measure of the spread within one trait, when one or more others have been "partialled out" or "held constant." It is used in determining the regression coefficient with which to "weight" each factor in the prediction.

Penalize: To subtract something from the score. Scoring two-answer questions Right-Wrong, penalizes for each error. It takes off two instead of one, from the total number correct.

Percentile: Each percentile point on the scale marks off one-hundredth of the group. A percentile score indicates the per cent of the total group who fell below the given score.

Plotting: Locating points on a graph.

Practical Research: Research which aims directly to improve an on-going activity. It grows out of a felt need in that practical concern, and must end with the improvement of that activity.

Practice Effect: After a person has gone through a test or operation, he often will do better or different the next time. This change in ability or attitude is known as the practice effect.

Predictive Index (P. I.): An interpretation of correlation in terms of the accuracy with which (by means of regression) scores in one trait can be predicted from knowledge of scores in the other. Where the P. I. is 1.00 the prediction is errorless, where it is zero the prediction is a pure guess. The P. I. is equal to $1.00 - K$.

Probable Error (P. E.): A measure of distance along the base line of a normal curve which includes half of the scores, if it be laid off on both sides of the mean. Better known as a unit for measuring reliability. Within one probable error either side of the obtained measure, there is a

1 to 1 chance that the true measures lies. Within 5 P. E.'s one can be practically certain that the true measure for a very large number of cases only a sample of whom have been tested, will fall. The probable error is always 0.6745 times the standard deviation, or standard error.

Product: The result obtained when figures are multiplied together.

Pure Research: Research which aims to find the true answer to a given question. It may appear to be very remote from practical concerns.

Quartiles, Q_1, Q_3: The points below which are found (Q_1) one-quarter of the scores, or (Q_3) three-quarters of the scores. To fall in the upper quartile, means to be within the best one-fourth of the group.

Q: May be called the semi-interquartile range. It is the average of the distance between Q_1 and Q_3. Or it may be said to be half the distance between Q_1 and Q_3. If laid off on either side of the median, then one-half of the cases would ordinarily be included. It is a measure of variability or spread. The larger it is, the more diverse and varied the group, on that particular test. In the normal curve Q = P. E.

Range: The distance from the lowest to the highest score. In careful work the distance from the lowest part of the lowest score to upper part of highest.

Rank, Ranking Questions: Ranking involves arranging in order. The best, truest, most desirable answer, is ranked one, the next two, etc.

Rank Distribution: An arrangement of scores which gives to each its place in the list, one being best, two next, etc.

Rate, Rating, Rater: In test making, answers that are rated may be assigned certain values, say from 1 to 5, or from 1 to 100. Word scales may be used instead of numbers, and the suggestion may be rated as excellent, good, fair, or poor, etc. In like fashion, people may be rated. The judgment and observation of the rater is the guide. The values that he assigns to the persons rated on certain traits constitute his ratings.

Ratio: The quotient when one number is divided by another and thus is stated in proportion to it.

Recognition Tests: See *Controlled Answer.*

Refinement: Elimination of errors, use of narrower units.

Regression: Technique for determining the most probable score in one trait from knowledge of score in the other. Based upon correlation between the two, and spread within each group. The regression coefficient is the weight by which a unit of distance in a given trait must be multiplied to make it correspond to a movement of one unit in the other trait.

Reliable, Reliability: A reliable test is one which yields consistently the same results under the same circumstances. Reliability is usually measured by the correlation between the first administration of a test

or rating scale, and the results secured by a second administration to the same group under similar circumstances and conditions.

Reliability is used in another sense, for any measurement, to indicate how much variation might be expected in the result if other samples were used. Any sample may give results which vary from what would have been found had everybody been tested. Within three standard deviations (or less than 5 probable errors) either side of a measure, it is practically certain that the true measure, which would be found from using all possible cases, falls. The S. D. and P. E. are thus measures of reliability in this sense. When they are small the measure is more reliable than in cases where they are larger. They are dependent on the number of cases. As the numbers grow larger, the measures of spread become smaller, hence the reliability is more encouraging. The true measure is more definitely located when its S. D. or P. E. is comparatively insignificant.

Retest: To test again, usually after an operation has been performed on the group.

Rho, ρ: A measure of correlation based upon rank distributions, rather than upon actual scores. Can be converted into r, the usual Pearson correlation coefficient, by means of Table XXIX.

Rotation: So arranging an experiment that every group in turn becomes subjected to every operation, that at any time every operation is in use in some group, and the practice effect is equated. One group starts with the first operation, another with the second, etc. Then the groups shift, each moving up one operation. This is continued in regular order.

Sample, Sampling, Fair Sample, Representative Sample: Most experimental work is based upon the theory that results from a small group are not far from what would be obtained if the total group could be measured. This small group which is taken as a representation of a larger group is called a sample. When certain conditions are met to avoid constant errors, the sample is said to be fair or representative. This occurs when numbers are drawn for a draft. The first 100 are a sample of all the thousands yet undrawn. The more that are drawn the better, fairer, and more representative the sampling is likely to become. In dealing with people, samples are often not chance. The group to be studied may be college students. They are not chance samples of all people. However, 25 students may be a fair sample of all those in college. If, of the 25, 20 are freshmen, then the sample is probably not fair to the total college group. The sample should be an approximate miniature of the total group being studied.

Sample Scale: A rating scale in which samples or illustrations are given to guide the judge in assigning values.

Score: The result of a measurement, whether of a test, or a count, a rating, etc.

Score Card: An instrument for rating which indicates the points to be observed, and often indicates the amount of weight to be given to each point.

Selection, Error of: Using as a fair sample a group which contains persons who are not really representative. Particularly operative in the case of people who answer questionnaires, or who voluntarily participate in an experiment. They are somehow different from the rest of the large group and conclusions based upon them are not valid for groups selected on other bases.

Self-correlation: The correlation obtained between one use of a test or rating scale, and the use a second time on the same group under the same conditions, Reliability.

Self-rating: A rating which a person makes of himself.

Sigma, σ. See *Standard Deviation.*

Significant Difference: A difference which is at least three times as large as the Standard Deviation of that difference.

Significant Factor: An element in the situation which makes a difference in the outcome. This must be judged in the light of the purpose of the experiment or investigation. An element that is significant for one experiment, might not be significant for another. Usually only the most significant ones are taken into account in equating groups, or measuring the strength of operating factors.

Single Group Methods: Experimental methods which can operate with only one group, or individual.

Skew: If the high point of a distribution curve is nearer either end than it is to the other, then it is said to be skewed. In a normal curve there is no skew, for the high bulge comes right in the middle. A skew is said to be positive if most of the scores are at the low end of the scale, with only a few stringing out toward the high end, but negative if the scores tend to pile up near the high end of the scale.

Spread: The tendency of scores to scatter out widely from the middle, Variation.

Standard Deviation, S. D., Sigma, σ: A measure of spread or variability, indicating a distance which, if laid off each side of the mean on a normal curve, would include about two-thirds of the cases. It is the best known and most used unit in which to measure spread. If the S. D. in one group is large, the group has a wide range, big scatter, much variation in the trait measured, when compared with another in which the S. D. is small. The S. D. is also used as a measure of reliability. If the entire group of possible responses were studied instead of just the sample taken, it is practically certain that the true result would fall

within three S. D.'s either side of the answer obtained from using a sample.

Standard Error: When the standard deviation is used in the latter sense of a measure of reliability it is often called the standard error.

Standardized Test: A test for which the reliability is known, validity is established, which is as objective as it can be made, and which has norms available so that new scores can be interpreted readily. A standardized test prescribes a definite carefully controlled procedure for giving and scoring.

Statistically Equated: Instead of getting groups actually equal by choice of individuals it is possible to measure the factors upon which one would like to see the groups equated, and by partial correlation to see what the result would have been, had they been actually equated upon those factors.

Step Interval: In a frequency distribution, the scores are sometimes grouped in step intervals. For example, it may be arranged to show how many scores between 70 and 74.9, how many between 75 and 79.9, how many between 80 and 84.9, etc. In such a case the step interval may be said to be 5. When scores are grouped together, the step interval is the range of the group.

Stimulus: A single something which is done, causing something else to happen, a causal factor.

Subjects: The people, objects, groups, institutions, etc., on which a test or experiment is tried. The people or objects which are rated, measured, or observed.

Sum: The result of adding together two or more figures. Often in formulae Σ is used to indicate a summation.

Survey Methods: Investigations which strive to gather up the results of experiments which have been performed in the natural course of life events. On the basis of accurate description of what has happened in certain groups, noting what characteristics go together, important experimental conclusions may be drawn.

Tabulation: Gathering together scores from individual papers or parts of papers, arranging them for statistical computation.

Technique: A definite way of doing something which has been carefully developed and which usually implies some skill for its use.

Test: A measure of characteristics, power, capacity, achievement, etc., which yields results which are dependent on what person measured would like to have it yield.

Time Test: A test in which speed is an important factor, and time is limited.

Trait: Any characteristic of a person, group, object or institution which can be defined or observed.

True-false Question: A new-type question consisting of a statement to which the subject must react by saying either that he believes it true and right, or partially or wholly wrong.

Two-answer Question: A controlled answer question in which there are only two possible responses. Besides the true-false, this general heading would include questions using right-wrong, good-poor, strong-weak, desirable-undesirable, like-dislike, cross-out or leave-standing, etc.

Unreliable: See *Reliable.*

Validity: That characteristic of a test or measurement, which shows that the test measures what it purports to measure. Evidence for validity is evidence that the test is well-named, that it does the thing people expect it to do.

Variability: Spread, range, scatter, tendency not to be close together.

Variables: Factors which might be important in accounting for differences between subjects or operating factors. If known and controlled they may be operating factors. If unknown and uncontrolled they may balance each other, or become errors.

Weight, Weighted, Weighting: The value assigned to a measure because of its importance in producing a certain result. Factors which make little difference should receive little weight, others which are very significant receive large weight. Weights may be assigned arbitrarily by shrewd guesses, or by some statistical device, such as the regression coefficient.

Word-phrase-answer Questions: Questions which test the ability of the subject to give the proper word, figure, or phrase.

INDEX

Accounting, method suggested, 225
Administration of tests, 122
Age norms, 153, 165
Aggressiveness, test for, 107
Aims of religious education, 67
Alienation, coefficient of, 208, 275
Army rating scale, 43
Association, tests using, 85, 100, 107
Attenuation, correction for, 209
Average deviation, 191

Bailor, predictive index, 208, 275
Biblical knowledge tests, 71, 86, 98, 99, 101
Bibliography, on statistics, 172; on tests, 71
Bogardus social distance test, 72
Bonser, party test, 109
Bridges, test of decision, 110
Brief test of religious education, 72
Brotmarkle comparison tests, 73
Brown prophecy formula, 161

Cady, 114
Case studies, 62
Case, true-false test in religious education, 97
Cattell-Fullerton theorem, 270
Caution, test of, 107, 120
Central tendency, measures of, 183
Certainty, makes ratings more reliable, 41
Chapman notives test, 97
Character education inquiry, 97, 110, 113, 114, 117; growth tests, 104; judgment, 35; reading, 35
Chassell parable test, 98
Cheating tests, 114–118
Cheerfulness, test of, 107
Choice of problem, 10
Church-school examination alpha, 74
Civic duty test, 108
Classification of test questions, 145, 148
Coefficient of alienation, 208, 275
Colgate emotional hygiene tests, 40, 75
Comparability of test scores, 165
Comparison of scores on different tests, in percentile units, 152; in SD units, 153, 194, 196
Completion questions, construction of, 133
Comprehensiveness of a test, 129, 167
Concentration, test of, 108
Conduct tests, 59, 106
Confidence, test of, 109
Conflict test. 100
Conformity, test of, 109
Consistency, test of, 105
Construction of test questions, 126,–150
Content of tests, selection of, 127
Contingency, coefficient of, 203
Controls, 16–17
Correlation, purpose of, 27, 197; computation of, 199–207; interpretation of, 207–209; self-correlation, 44; partial correlation, 28; correction for attenuation, 209
Cost of research, 12, 237
Courtesy, test of, 109
Criteria for good tests, 166–169
Criterion score, use in building scoring schemes, 153–158
Critical score, 212
Curriculum, suggested experiments in, 264–266

Decile, 188
Decision, test of, 110
Definition of problem, 10
Delinquency, test of, 110
Dependability, test of, 114
Depression, test of, 107
Detroit primary tests, 69
Deutsch, test of conformity, 109
Directions for tests, construction of, 148–150
Discussion of test results, 123
Displacement, when predicting from correlations, 208, 276

Distribution, frequency, 178; order, 176; rank, 176
Downey will temperament test, 76, 121

Emotional history record, 78, 100; tests, 75, 78, 85, 87, 89, 95, 112–113
Equated group method, 17–24
Error of prediction of correlation by regression, 207
Essay examination, compared with new type, 128
Eta (η), 199
Ethical tests, 72, 73, 80, 85, 93, 97, 99, 101, 103, 110
Ethics of testing, 59
Experimentation, problems not answered by, 3
Experiments, suggested, 256–269

Fahs Biblical test, 99
Fair-mindedness, test of, 92
Faking test results, 58, v,
Fernald achievement capacity test, 79; ethical discrimination test, 80; meritorious acts and ambitions, 99
Form of test question, types of, importance of, 70
Frequency distribution, 178
Freyd's occupational interest blanks, 81
Furfey, study of ratings, 37

Gates, test of social perception, 120
Giles Sunday school examination A, 82
Glossary, 279–291
Good manners test, 28
Grade norms, 153, 165
Grades, 189
Graph, forms of, 248; rules for making, 249
Graphic distribution, 178; rating scale, 39, 41
Group loyalty, test of, 113

Haggerty intelligence tests, 69
Halo effect, 40
Hart, personnel assayer, 83; test of social attitudes and interests, 83
Hartshorne, 38, 97, 99, 114
Helpful behavior, 113
Herring-Binet test, 68
Home environment tests, 114
Homogenity, measures of, 191; effect on reliability of test, 163
Honest confession, 117
Honesty, tests of, 114–118
Houck, study of prohibition attitudes, 17
Human-ladder scale, 43
Humor, test of, 118

Illinois intelligence tests, 69
Imagination test of, 118
Indiana survey, 38
Individuals, application of experiments to, 31
Institute of educational research tests of intelligence, 68–69
Intellectual immoralities, 252
Intelligence, tests of, 68; validity of tests, 51
Interest analysis, 84
Interests tests, 81, 88, 91, 102, 119
International attitudes tests, 102
Interviews, 61

Kent Rosanoff free association test, 85
Koh's ethical discrimination test, 85, 110

Landis, 41
Laslett, association tests, 100
Laycock test of Biblical information, 86
Length of tests, effect on reliability, 162
Lentz, studies in character tests, 49, 110–111
Life-situation tests, 106
Linearity of regression, 199
Lundholm emotional cross-out tests, 87

Man-to-man scale, 43
Manual, preparation of, 160
Marks, 189
Matching questions, 144
May, Mark, 38, 67, 97, 114, 117, 120
McCaskill, test of group loyalty, 113
McGeoch, imagination test, 119
Mean, 183
Measures, really universal, 34; refined, 35
Median, 184
Mental hygiene tests, 40, 75, 78, 95, 100; tests. *See* Intelligence Tests
Methods of instruction, effect of tests on, 166–167
Miner's analysis of work interest blank, 88
Mode, 186

Moore, test of aggressiveness, 107
Multi-mental tests, 69
Multiple-choice questions, construction of, 135; tests of religious ideas, 89
Multiple correlation, 222

National intelligence tests, 69
Negativism, 119
New type examinations, compared with essays, 128
Normal frequency curve, 173, 181
Norms, basis for in religious education, 153, 165

Objectivity, provision for in tests, 128, 160
Order distribution, 176
Orr, good manners test, 28, 101
Otis, self-administering intelligence tests, 69; tests of suggestibility, 121

Pacifism-militarism experiment, 102
Pairing pupils of equated groups, 19; questions, 144
Paper and pencil tests, 57
Parable interpretation test,
Partial correlation, 213; two variable, 214; illustrative problem, 224–237
Partial regression, regression coefficient, 218; standard deviations, 219
Pearson's coefficient of correlation, 197, 203
Percentage equivalent for correlation, 209, 277
Percentile, units for scoring tests, 152; computation of percentiles, 188
Persistence, test of, 120
Physiological measures, 112
Picture-preference test, 103
Pintner-Cunningham tests, 69
Pintner Patterson performance test, 69
Porter, advance Bible test, 101; student opinion on war, 102
Posters, use in presenting results, 248
Practical certainty, 238; research, steps in, 5; choice of problems for, 6
Predictive index, 208, 275
Prejudice tests, 72, 92, 105
Pressey X-O tests, 89, 110
Probable error, 237
Problems, suggested for experimental investigation, 256, 269
Prohibition, experiment in influencing attitudes towards, 17
Prophecy formula, 161
Pure research, 5
Purpose of an experiment, 10

Q, 196
Quartiles, 186
Questionnaires, use of, 55; distinquished from tests, 58

Racial attitudes test, 91
Range, 191
Rank distribution, 176; method for computing correlation, 199, 203
Ranking questions, construction of, 142; directions for, 142, 149
Rating scales, construction of, 35; value of, 43, 53
Ratings, methods of improving, 35–43
Raubenheimer, 100, 102, 112, 114
Ream's social relations test, 91
Recklessness, test of, 120
Reconstruction of ethical standards needed, 67, 147–148
Regression coefficient, 211
Relationship, measures of, 197–237
Reliability, meaning of, 44, 237; tests of, 44; of character ratings, 44; of correlation, 240; difference, 241; mean, 237; median, 238; Q, 239; regression, 240; standard deviation, 237; of tests, computation of, 160; influence of variation in group, 163; influence of length of test, 162
Religious ideas tests, 72, 89, 94, 97, 98
Report, preparation of, 247
Report blanks with tests, 122
Rho (ρ) method of correlation, computation of, 199–203; transmutation to "r", 274
Rotation method, 24
Rugg, study of ratings, 37, 45

Sample scale, 42, 270
Schwesinger, ethical vocabulary study, 28, 102
Score cards, construction of for church plants, 38
Scoring scheme, development of, 150
Scott rating scale, 43
S. D. *See* Standard Deviation

S. D. units in terms of proportion of area curve above them, 278
Selection of test content, 127
Self-esteem test, 100
Self-ordinary-ideal rating scale, 100
Sigma. See Standard Deviation
Sims, scale for social-economic status, 114
Single group method, description of, 13; advantages and limitations of, 15
Skew, negative or positive, 181
Sociability, 120
Social distance tests, 72
Social-economic status, scale for, 114
Social perception, test of, 120
Socio-ethical vocabulary test, 28
Spearman-Brown prophecy formula, 161
Spread, measures of, 191, 197
Squares and square roots, tables of, 273
Standard deviation, computation of, 193; meaning of, 193, 194; use as unit, 153, 194–196; use in reliability measure, 237–243, 237; error, 238
Standardization, 125; of tests, 159
Stanford-Binet test, 68
Statistics, need for, 171
Step interval, of frequency distribution, 178
Studiousness, test of, 120
Sturges, 15
Suggestibility, tests of, 121
Survey of public opinion on some religious and economic questions, 92
Surveys, 26
Symonds, test of studiousness, 41, 58

T scale, 152, 196
Terman group tests of intelligence, 69
Tests, criteria for good, 166–169
Thorndike examinations, 68
Travis, test of personality traits, 103
Trow, test of confidence, 109
True false questions, construction of, 129; directions for, 130, 149; scoring of, 151
Trustworthiness, tests of, 114, 118
Truthfulness, test of, 114–118
Two-answer questions, construction of, 131
Types of experiments, 10–29

Unfinished story test, 99
Upton-Chassell citizenship scale, 49, 95

Validity, meaning of, 47,164; methods of study, 47, 164
Van Wagenen, history tests, 103
Variability, measures of, 191 197; effect on reliability, 163
Vocabulary tests, 28, 102
Vocational tests, 68
Voelker tests, 52, 61, 108, 110, 114, 122

War, test of attitudes towards, 102
Washburn, tests of cheerfulness, 107
Weighting, by regression coefficient, 221; of score cards, 39; of test elements, 158
Whipple, test of fidelity, 109, 121
Whitley test, 71
Whittier scales, 114
Will-power test, 79
Woodrow-Picture-preference test, 103
Woodworth-Mathews personal data sheets, 95
Word-phrase answer questions, construction of, 134

Year books, 50, 54
Y M C A religious education tests, 104, 119

268
W333e

80720